Istat
Istituto Nazionale di Statistica

L'INDAGINE EU-SILC: INNOVAZIONI NELLA METODOLOGIA DI RILEVAZIONE E DI STIMA

METODI · LETTURE STATISTICHE ·

L'INDAGINE EU-SILC: INNOVAZIONI NELLA METODOLOGIA DI RILEVAZIONE E DI STIMA

EDIZIONE 2021

Attività editoriali: Nadia Mignolli (coordinamento), Marzia Albanesi, Patrizia Balzano e Alessandro Franzò.

Copertina: Maurizio Bonsignori.

ISBN 978-88-458-2045-8

STREETLIB

INDICE

Pag.

Introduzione 9

PARTE PRIMA

1. Dal disegno di indagine alla rilevazione sul campo 15
1.1 Il disegno campionario 15
1.2 Le unità di rilevazione 17
1.3 Le diverse tecniche di rilevazione utilizzate 18
1.4 Le misure per la prevenzione e controllo degli errori di rilevazione 19
1.5 Sviluppo e test del questionario 20

2. La sperimentazione della tecnica CATI e la progettazione della rilevazione con tecnica mista CAPI/CATI 21
2.1 Introduzione 21
2.2 Il contesto italiano: il passaggio da CAPI a CATI 22
2.3 La sperimentazione CATI 23
2.3.1 Organizzazione della sperimentazione e selezione del campione 23
2.3.2 Riprogettazione del questionario 24
2.3.3 Test del questionario 26
2.3.4 Formazione dei rilevatori 27
2.3.5 Rilevazione, sistema di indicatori di monitoraggio, monitoraggio di sala 28
2.3.6 I risultati della sperimentazione 31
2.4 Rilevazione con tecnica mista CAPI/CATI 32
2.4.1 La progettazione della tecnica mista CAPI/CATI 32
2.4.2 La tecnica mista CAPI/CATI: i risultati della rilevazione dal 2016 al 2018 34
2.5 Conclusioni 36

PARTE SECONDA

3. I metodi e le tecniche per il controllo, la correzione, l'imputazione e la validazione dei dati 41
3.1 Il controllo e la correzione dei dati 41
3.2 Le fasi del processo di controllo e correzione dei dati 42

Pag.

4. SAGE: Il Sistema automatico di generazione delle regole di incompatibilità nelle variabili del questionario 49
4.1 Introduzione 49
4.2 Alcune definizioni e aspetti teorici 50
4.3 I metadati del questionario 51
4.4 Il metodo di correzione probabilistica adottato in EU-SILC 53
4.5 La generazione automatica degli *edit* 54
4.6 Un esempio di generazione automatica degli *edit* formali (EU-SILC 2016) 55
4.7 *Edit* sostanziali e conseguenze sulle partizioni generate automaticamente 57
4.8 Conclusioni 60

5. FRI: Una nuova procedura per l'imputazione da donatore 61
5.1 Introduzione 61
5.2 Cenni metodologici 61
5.3 La procedura di imputazione totale da donatore 62
5.4 Gli output della procedura 64
5.5 Un'applicazione ai dati dell'indagine EU-SILC 65
5.6 Conclusioni 68

6. La stima delle tasse e dei contributi sociali 69
6.1 Introduzione 69
6.2 Alcune specificità della metodologia di stima dell'imposizione fiscale e contributiva 69
6.3 La parametrizzazione dei contributi sociali nel modello SM2-EU-SILC 71
6.4 Le imposte micro-simulate 74
6.5 Conclusioni 79

7. Il processo di validazione con fonti esterne a supporto del trattamento dei dati 81
7.1 Introduzione 81
7.2 La validazione dei redditi pensionistici 82
7.3 La validazione dei redditi da lavoro dipendente 84
7.4 La validazione dei trasferimenti non pensionistici 85
7.5 Conclusioni 86

PARTE TERZA

8. Le fonti amministrative per l'integrazione dei dati d'indagine 89
8.1 Introduzione 89
8.2 Excursus storico del processo d'integrazione di microdati in EU-SILC 90
8.3 Presentazione dei contributi della parte terza 91

9. L'impatto dei dati amministrativi sulle stime finali dei redditi 93
9.1 Introduzione 93
9.2 Dati e metodi 94

Pag.

9.3 Risultati: database IT-SILC Solo Campione vs Integrato Finale 94
9.3.1 Costruzione del database Solo Campione 94
9.3.2 Distribuzione del reddito totale familiare netto 95
9.3.3 Rischio di povertà 96
9.3.4 Principali componenti del reddito individuale e loro profili 97
9.4 Risultati: database IT-SILC Solo Fisco vs Integrato Finale 98
9.4.1 Costruzione del database Solo Fisco 98
9.4.2 Principali componenti del reddito individuale e loro profili 99
9.4.3 Quali delle componenti di reddito di fonte campionaria potrebbero essere sostituite con dati di fonte amministrativa? 101
9.5 Conclusioni 103

10. La stima della cassa integrazione anticipata dai datori di lavoro 105
10.1 Introduzione 105
10.2 Le tipologie di cassa integrazione: caratteristiche e riferimenti normativi 106
10.3 Ricognizione delle fonti dati e strategie di rilevazione della cassa integrazione anticipata 107
10.4 Trattamento preliminare dei dati di archivi amministrativi per l'uso della variabile differenza/accredito 109
10.5 Metodo di calcolo 113
10.6 Risultati dell'implementazione 116

11. Un'analisi esplorativa dell'archivio Inps su certificazioni telematiche di malattia 117
11.1 Introduzione 117
11.2 Malattia: identificazione dell'unità di analisi, aspetti concettuali e classificazioni 118
11.3 Malattia: le fonti del dato statistico 120
11.3.1 Certificazione telematica della malattia dei dipendenti pubblici 121
11.3.2 Certificazione telematica della malattia dei dipendenti privati 121
11.3.3 Archivio Inps prestazioni dirette non pensionistiche 122
11.3.4 Archivio Inps Emens delle differenze di accredito 122
11.3.5 Archivio Inps Emens dei conguagli 122
11.3.6 Casellario Inps delle posizioni previdenziali attive 122
11.4 Malattia: consolidamento archivi, criteri di calcolo e disegno multifonte 123
11.4.1 La stima indiretta della prestazione di malattia e l'elemento base: periodo 123
11.4.2 La stima diretta della prestazione di malattia: importo erogato 128
11.5 Analisi delle incoerenze, riconciliazione tra fonti e validazione esterna 132
11.6 Conclusioni 135

12. L'utilizzo della Banca Dati Reddituale del MEF per la calibrazione del campione 137
12.1 Introduzione 137
12.2 La scelta delle variabili della BDR 138

Pag.

12.3 Inserimento delle variabili della BDR nel campione EU-SILC 139
12.4 Confronto delle stime EU-SILC con i totali noti della BDR 140
12.5 La procedura di costruzione dei coefficienti di riporto all'universo e le variabili BDR 144
12.5.1 Le variabili BDR nella correzione per la mancata risposta totale 144
12.5.2 Le variabili BDR nella calibrazione del campione 146
12.6 L'impatto dei nuovi coefficienti di riporto all'universo sulle stime finali 146
12.7 Conclusioni 150

PARTE QUARTA

13. Il trattamento e l'utilizzo dei dati longitudinali 153
13.1 Introduzione 153
13.2 Unità di rilevazione longitudinale e regole di inseguimento 154
13.3 Principali criticità della rilevazione longitudinale 155

14. La trasformazione dei file provenienti dalla rilevazione alla base dei processi di trattamento dei dati dell'indagine 157
14.1 Introduzione 157
14.2 L'organizzazione dei dati 158
14.3 Il software di trasformazione dei dati 159
14.4 L'algoritmo di creazione degli identificativi individuali 160
14.5 Il *software* di trasformazione dei dati 163
14.6 Conclusioni 164

15. L'impatto della rilevazione proattiva delle informazioni sulla componente longitudinale 165
15.1 Questionario elettronico e uso delle domande a conferma nelle indagini panel 165
15.2 Questionario elettronico, domande a conferma e strategie di utilizzo delle informazioni rilevate nelle interviste precedenti 166
15.3 Effetti delle domande a conferma sulle transizioni rilevate tra coppie di anni 168
15.3.1 Questionario familiare 169
15.3.2 Questionario individuale 172
15.4 Conclusioni 174
Appendice 176

16. Il trattamento dei dati longitudinali 181
16.1 Introduzione 181
16.2 Strategia di correzione longitudinale in EU-SILC 182
16.3 Correzione "all'indietro" 183
16.4 Correzione "in avanti" 186
16.5 Correzione basata sulla "prevalenza" 187
16.6 Conclusioni 189

Pag.

17. L'utilizzo dei dati in ottica longitudinale: potenzialità del disegno campionario a gruppi rotazionali 191
17.1 Introduzione 191
17.2 Campioni e popolazioni longitudinali 191
17.3 Sistema di pesi longitudinali e strategia di stima 195
17.3.1 Pesi longitudinali per panel e anno di osservazione (RB060) 195
17.3.2 Pesi per la stima di indicatori su unità compresenti (RB062, RB063 e RB064) 196
17.4 Conclusioni 199

Riferimenti bibliografici 201

INTRODUZIONE[1]

L'indagine sul reddito e le condizioni di vita delle famiglie nasce all'interno di un più ampio progetto denominato "*Statistics on Income and Living conditions*" (EU-SILC) deliberato dal Parlamento europeo e coordinato da Eurostat[2]. Tale progetto risponde alla sempre più ampia e dettagliata richiesta di informazione statistica su argomenti come redditi, povertà, esclusione sociale, deprivazione, qualità della vita.

La necessità di un ampio bacino di indicatori su queste tematiche, nonché la profonda importanza di una loro armonizzazione a livello comunitario per permettere gli opportuni confronti, persegue gli obiettivi che l'Unione europea si è impegnata a raggiungere nel Consiglio di Lisbona (marzo 2000) e con la Dichiarazione di Laeken (dicembre 2001), ovvero un'economia competitiva e dinamica, basata sulla conoscenza e sulla sostenibilità e rivolta al miglioramento della coesione sociale.

Grazie a questo progetto, Eurostat e gli Istituti nazionali di statistica europei[3] mettono a disposizione degli studiosi, delle autorità di politica economica e dei cittadini una serie di dati sulle condizioni di vita delle famiglie, cioè informazioni a livello familiare e individuale sui redditi e su altre dimensioni che determinano il benessere materiale e, più in generale, la qualità della vita.

Sulla base di questi dati, la UE calcola gli indicatori ufficiali per la definizione e il monitoraggio degli obiettivi di politica sociale per stimolare lo sviluppo e l'occupazione nell'UE, nel contesto della Strategia Europa 2020, che punta a far uscire dal rischio di povertà o di esclusione sociale almeno 20 milioni di persone entro il 2020.

Questo obiettivo è monitorato attraverso l'indicatore di povertà o esclusione sociale, che misura la quota di persone che sperimentano almeno una condizione tra: rischio di povertà, grave deprivazione materiale, molto bassa intensità lavorativa nella famiglia.

L'indagine offre un ampio spettro di informazioni sulle condizioni di vita delle famiglie, a livello sia familiare sia individuale, di carattere sia monetario sia non monetario. In particolare, consente di valutare:

- la **povertà**, basata sui redditi netti (rilevati attraverso l'intervista diretta e/o ricostruita attraverso integrazione con dati di fonte amministrativa);
- la **deprivazione materiale** (incapacità, per scarsità di risorse economiche, di accedere a beni e servizi essenziali);
- le **condizioni abitative** (problemi strutturali, sovraffollamento);

1 Il volume è stato curato da Paolo Consolini, Lucia Coppola, Isabella Siciliani e Silvano Vitaletti. Alla sua redazione hanno contribuito: Giovanni Battista Arcieri, Maria Cirelli, Paolo Consolini, Lucia Coppola, Clodia Delle Fratte, Gabriella Donatiello, Alessandro Fratoni, Doriana Frattarola, Stefano Gerosa, Francesca Lariccia, Daniela Lo Castro, Pierpaolo Massoli, Mattia Spaziani e Silvano Vitaletti.

2 L'indagine EU-SILC è stata normata a livello europeo fino al 31 dicembre 2020 dal regolamento quadro (Framework regulation) n. 1177/2003 (emendato dal regolamento n. 1553/2005 Enlargement and derogations e dal regolamento n. 1791/2006 Enlargement) e dai successivi regolamenti di esecuzione. A decorrere dal 31 dicembre 2020 il regolamento n.1177/2003 è stato abrogato dal regolamento 2019/1700 del Parlamento europeo e del Consiglio del 10 ottobre 2019 che istituisce un quadro comune per le statistiche europee sulle persone e sulle famiglie, basate su dati a livello individuale ottenuti su campioni (IESS, Integrated European Social Statistics regulation) cui faranno seguito i relativi regolamenti di esecuzione. Per il quadro completo e aggiornato della legislazione comunitaria su EU-SILC si veda la pagina http://ec.europa.eu/eurostat/web/income-and-living-conditions/legislation.

3 Allo stato attuale, il progetto EU-SILC è stato implementato dai 27 Paesi UE e da Gran Bretagna, Islanda, Norvegia, Svizzera, Macedonia del Nord, Serbia, Turchia e Montenegro. È in corso di implementazione in Albania e in Bosnia-Erzegovina.

- il **disagio economico** soggettivo (giudizio sulla capacità di arrivare senza difficoltà alla fine del mese, onerosità delle spese necessarie alla famiglia).

Il fulcro dell'indagine, ovvero il **reddito netto** nelle sue principali componenti, è raccolto principalmente a livello individuale: la somma delle varie tipologie di reddito permette di ricostruire il reddito familiare netto disponibile, funzionale alla costruzione della soglia di povertà, al di sotto della quale la famiglia (e tutti i suoi componenti) viene classificata a rischio di povertà.

Oltre ai redditi netti vengono ricavati i **redditi lordi**, comprensivi delle imposte e dei contributi sociali, stimati mediante un modello di micro-simulazione e l'integrazione dei dati campionari e amministrativi.

L'indagine raccoglie inoltre informazioni dettagliate, a livello individuale, sul **livello di istruzione**, sulle **condizioni di salute** e l'**accesso ad alcuni servizi sanitari**, sulla **situazione occupazionale** (sia oggettiva, sia auto-dichiarata), sulle **caratteristiche del lavoro**, attuale o ultimo svolto.

Una sezione dell'indagine è infine dedicata al ricorso delle famiglie a strutture formali, informali e scolastiche, a pagamento o gratuite, per **l'affidamento dei bambini** di età fino ai 12 anni.

L'indagine produce una serie di dati con riferimenti temporali diversi: le caratteristiche familiari e individuali fanno riferimento al momento dell'intervista; le spese sostenute dalle famiglie per l'abitazione si riferiscono agli ultimi 12 mesi; gli indicatori di deprivazione e di benessere al momento dell'intervista; tutti i dati inerenti il reddito all'anno solare precedente quello di rilevazione.

L'indagine, che ha cadenza annuale, fornisce anche dati **longitudinali** (fino a quattro osservazioni ripetute che nel 2020 sono diventate cinque per poi passare a sei nel 2021 quando il panel d'indagine diventa sessennale[4]), che permettono di studiare le dinamiche dei fenomeni rilevati, con particolare attenzione a quella della povertà.

La principale **unità di analisi** è costituita dalla **famiglia di fatto**, definita come l'insieme delle "persone che vivono abitualmente attualmente presenti o temporaneamente assenti nella stessa abitazione, legate da vincoli di parentela, affinità, adozione, tutela, affetto o amicizia". Tale definizione, conforme agli standard adottati nelle indagini Istat sulle famiglie, viene adottata in deroga alla definizione di famiglia dettata dal regolamento europeo, essendone considerata una adeguata *proxy*[5].

Oltre alla famiglia di fatto, costituiscono specifiche unità di analisi (per le quali vengono ottenute stime riportate alla popolazione di riferimento) anche gli stessi individui appartenenti alle famiglie di fatto; gli individui di almeno 16 anni di età, cui è rivolta l'intervista personale (fino al 2010 rivolta anche ai 15enni); i bambini fino ai 12 anni di età, per i quali vengono rilevate informazioni relative al *child care* (per il tramite di un adulto della famiglia).

Il presente volume intende offrire una descrizione quanto più possibile esaustiva delle differenti fasi del processo di indagine, sia richiamando sinteticamente ciò che in modo più approfondito è già stato descritto in altre pubblicazioni metodologiche, sia mettendo

4 Il nuovo Regolamento quadro 2019/1700, applicato a partire dal primo gennaio 2021, stabilisce, tra le altre norme, i requisiti di tempestività e di precisione per i dati dell'indagine EU-SILC. Per ottemperare a questi requisiti, in Italia il disegno campionario è stato rafforzato a partire dal 2020, ampliando il campione (che supererà le 30mila famiglie) e portando a sei anni la lunghezza del panel al fine di permettere di irrobustire l'analisi longitudinale.

5 Il nuovo Regolamento quadro 2019/1700 definisce la famiglia "una persona che vive da sola o un gruppo di persone che vivono insieme e che provvede o provvedono autonomamente ai prodotti di prima necessità". È in corso quindi un confronto sulle modalità di implementazione di questa nuova definizione nei diversi Paesi europei e in particolare in Italia.

in risalto le innovazioni di metodo e di tecnica introdotte negli ultimi anni, meritevoli a nostro avviso di una condivisione più ampia rispetto a quella attuale.

Il volume può essere idealmente suddiviso in quattro parti, sequenziali ma fortemente interconnesse le une con le altre.

Nei primi due capitoli si articola la parte relativa alla rilevazione diretta dei dati presso le famiglie campione, con una disamina delle diverse tecniche di rilevazione che nel tempo si sono avvicendate.

Successivamente, nei capitoli che vanno dal terzo al settimo, vengono descritti i metodi utilizzati nelle fasi di trattamento dei dati raccolti, partendo dalla fase di controllo e correzione per poi passare alla fase di imputazione e di microsimulazione, per concludere con la fase di validazione.

Nella terza parte (che va dal capitolo 8 al capitolo 12) viene dato risalto all'utilizzo dei dati amministrativi nel processo di indagine, motivato sia dall'esigenza di ridurre quanto più possibile il carico sui rispondenti, sia dall'opportunità di conseguire un miglioramento della qualità dell'affidabilità dei risultati ottenuti, nonché di offrire una maggiore granularità delle informazioni statistiche.

Infine, nella quarta e ultima parte, che abbraccia i capitoli dal 13 al 17, vengono approfonditi alcuni tra gli aspetti più qualificanti dell'indagine EU-SILC e cioè quelli relativi alla componente longitudinale dell'indagine, sia per ciò che riguarda le specifiche fasi di trattamento dei dati, sia per ciò che riguarda i metodi più appropriati per il loro utilizzo nelle analisi dei fenomeni oggetto di studio.

In ciascuna delle quattro parti in cui si divide il volume viene prima offerta, in un capitolo introduttivo, una sintesi di quello che può essere considerato lo stato dell'arte in merito agli argomenti della sezione. Più in dettaglio, nei capitoli successivi, vengono esaminate le innovazioni che si sono radicate nel processo di indagine o che sono ancora allo stadio di sperimentazione prima di passare alla produzione corrente.

PARTE PRIMA

1. CAMPO DI OSSERVAZIONE E UNITÀ DI ANALISI[1]

1.1 Il disegno campionario

La progettazione del disegno campionario è stata guidata dai requisiti stabiliti dai regolamenti europei relativamente ai diversi aspetti della rilevazione, ovvero: la tipologia di parametri che l'indagine deve produrre, la cadenza e il periodo di riferimento dei quesiti, la dimensione minima campionaria utile a raggiungere le stime obiettivo di tipo trasversale e longitudinale.

Il disegno adottato è un **campionamento casuale composito a due stadi con stratificazione delle unità di I stadio**. La stratificazione dei comuni per dimensione demografica (all'interno di ciascuna regione) definisce sia strati formati da un solo comune auto-rappresentativo (AR), sia strati formati da più comuni (NAR). Per gli strati AR il campionamento è a un solo stadio con selezione casuale delle famiglie campione con metodo sistematico. Per gli strati NAR vengono selezionati casualmente 4 comuni campione (o soltanto 2 negli strati formati soltanto da 2 o 3 comuni) con probabilità di inclusione proporzionale alla dimensione demografica dei comuni stessi; le famiglie vengono selezionate all'interno dei comuni campione come nel caso AR.

Il disegno di indagine deve integrare una componente trasversale e una componente longitudinale. Per l'Italia, il campione relativo a ogni occasione d'indagine è un **panel ruotato** costituito da quattro gruppi rotazionali (ognuno di dimensione pari a un quarto della numerosità campionaria complessiva teorica). Ogni gruppo rimane nel campione per quattro anni consecutivi e ogni anno il campione si rinnova con l'entrata di un nuovo gruppo (Figura 1).

Fino al 2014 i gruppi rotazionali venivano associati ai comuni campione: in ogni comune le nuove famiglie da intervistare venivano estratte ogni 4 anni (ogni 2 anni per i comuni appartenenti a strati con solo 2 comuni campione, ogni anno per i comuni AR). Ogni anno, quindi, nei comuni appartenenti agli strati NAR con 4 comuni campione, venivano intervistate soltanto famiglie di 1ª, 2ª, 3ª o 4ª *wave* (oppure di 1ª e 3ª o 2ª e 4ª negli altri comuni NAR). Questo per rendere maggiormente sostenibile, da parte dei comuni più piccoli, l'estrazione delle famiglie dai propri registri. Dal 2012 tale esigenza è venuta meno, grazie alla disponibilità in Istat delle Liste Anagrafiche Comunali (LAC), che ha permesso di centralizzare l'estrazione delle famiglie-campione.

A partire dal 2015, invece, la rotazione del campione viene associata direttamente alle famiglie campione: in ogni comune viene estratta, ogni anno, una quota di nuove famiglie da intervistare. In questo modo, in ogni comune e in ogni anno di indagine, vengono intervistate sia famiglie di 1ª *wave*, sia famiglie di 2ª, 3ª e 4ª *wave*.

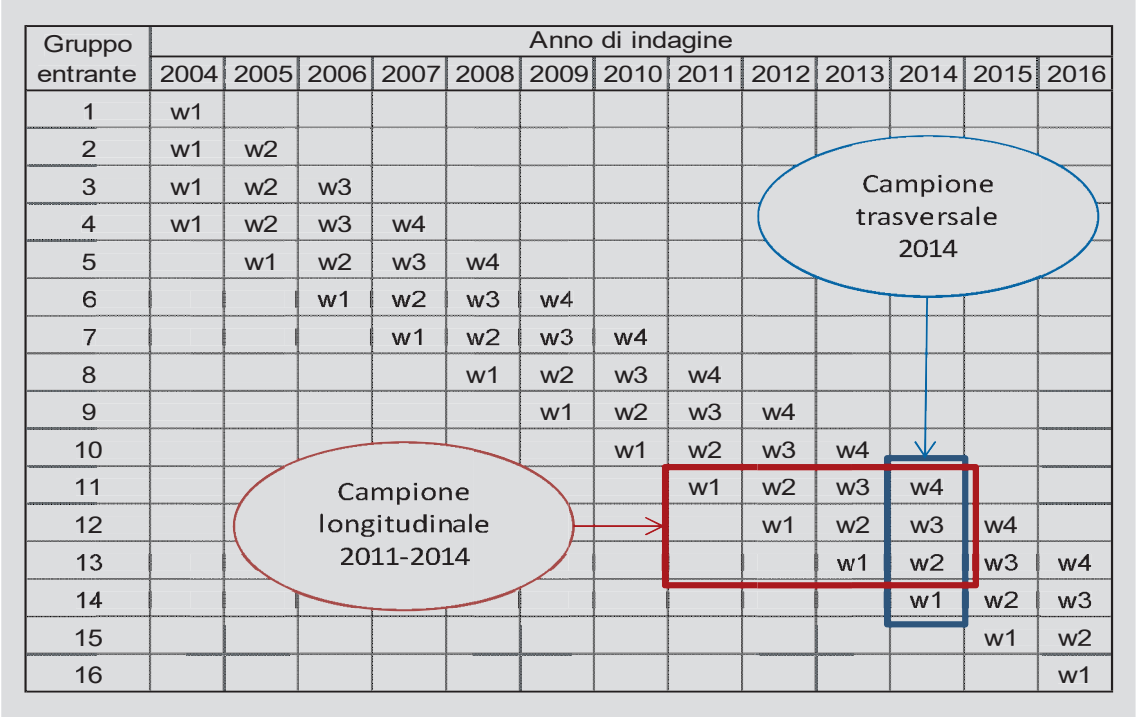
Dal 2015, inoltre, si è deciso di ridurre il numero complessivo dei comuni campione (originariamente 762) per riuscire a conseguire una migliore copertura del territorio da parte delle reti di rilevazione private[2] alle quali è stata affidata l'indagine. Sono quindi stati esclusi i comuni per i quali era prevista l'estrazione delle nuove famiglie nel 2015 (solo per

1 Il capitolo è stato redatto da Silvano Vitaletti.

2 A partire dall'edizione 2011 dell'indagine. Negli anni dal 2004 al 2010 la rilevazione è stata condotta tramite la rete di rilevazione comunale.

i 148 strati NAR con 4 comuni campione). In questo modo il numero di comuni campione in questi strati è passato a 3 e il totale dei comuni campione si è ridotto a 614.

Figura 1.1 - Schema rotazionale della rilevazione Eu-Silc



Oltre ai comuni del campione base, in ogni anno di indagine viene attivata la rilevazione anche in altri comuni per effetto dei **trasferimenti** delle famiglie o di singoli individui sul territorio nazionale. Gli individui che al momento della prima intervista avevano compiuto almeno 14 anni di età (**individui campione**) vengono infatti "**inseguiti**" nelle *wave* successive. Nel caso il trasferimento abbia riguardato soltanto alcuni componenti della famiglia, la nuova famiglia in cui questi si sono trasferiti viene interamente inclusa nel campione e intervistata.

Per ciò che riguarda la **numerosità del campione**, ai fini della comparabilità delle stime in termini di precisione tra i diversi paesi europei che partecipano all'indagine EU-SILC, Eurostat impone un minimo che tiene conto dell'effetto del disegno di campionamento (DEFF) sull'errore campionario associato alla stima della percentuale di individui a rischio di povertà. In pratica, a partire dalla numerosità del campione teorico di famiglie (*actual sample size*), si deve considerare la riduzione per effetto della mancata risposta totale (*achieved sample size*) e un'ulteriore riduzione per effetto del disegno di campionamento (*effective sample size=achieved sample size/DEFF*).

Nel caso dell'Italia, la numerosità effettiva minima del campione fissata dal regolamento europeo, definita separatamente per la componente trasversale e per quella longitudinale dell'indagine, è illustrata nella Tavola 1 mentre le numerosità effettive conseguite, facendo il caso del campione trasversale 2014 e del campione longitudinale 2013-2014, vengono mostrate nella Tavola 2.

Come si vede, la numerosità campionaria del campione nazionale eccede i requisiti minimi stabiliti da Eurostat. Ciò al fine di garantire le esigenze di informazione per livelli territoriali sub-nazionali (regioni e ripartizioni geografiche) di specifico interesse nazionale.

Tavola 1.1 - Numerosità minima del campione EU-SILC in base al regolamento europeo

Campione	Famiglie	Individui di 16 anni e più
Trasversale	7.250	15.500
Longitudinale (a)	5.500	11.750

(a) Per ogni coppia di anni consecutivi di indagine

Tavola 1.2 - Numerosità conseguita ed effettiva nel campione EU-SILC

Campione	DEFF	Famiglie		Individui di 16 anni e più	
		Numerosità conseguita	Numerosità effettiva	Numerosità conseguita	Numerosità effettiva
Trasversale 2014	1,61406	19.663	12.182	40.280	24.956
Longitudinale 2013-2014	1,40434	13.412	9.550	**27.732**	**19.747**

Fino al 2010, per garantire le numerosità fissate tenendo conto dell'effetto della mancata risposta (la rilevazione non prevede la sostituzione delle famiglie non rispondenti), la numerosità del campione entrante (1ª *wave*) è stata fissata a circa 8.000 famiglie.

Nel 2011, con il passaggio alla tecnica CAPI e in previsione di un minore tasso di risposta, la numerosità del campione entrante è stata portata a circa 10.000 famiglie.

Dal 2016, con l'ingresso a regime della tecnica di rilevazione mista CAPI-CATI, la numerosità del campione entrante è stata portata a circa 14.000 famiglie per compensare sia il maggiore impatto previsto della mancata risposta totale per le famiglie da intervistare telefonicamente, sia il progressivo depauperamento del campione che ha avuto luogo soprattutto tra il 2014 e il 2015.

1.2 Le unità di rilevazione

Le unità di rilevazione sono costituite dalle famiglie residenti in Italia, ovvero che risultano iscritte nelle anagrafi comunali. Fino al 2011 l'estrazione della lista di famiglie da intervistare veniva demandata ai singoli comuni. Successivamente, l'estrazione delle famiglie è stata centralizzata utilizzando le Liste Anagrafiche Comunali (LAC) entrate nella disponibilità dell'Istat.

Anche la definizione operativa di eleggibilità delle famiglie, necessaria per stabilire una corrispondenza tra l'unità di rilevazione e l'unità di analisi (la famiglia di fatto), ha subito nel tempo alcune modifiche.

Per le **famiglie del campione entrante**, fino al 2015, la famiglia anagrafica veniva considerata eleggibile se al suo indirizzo anagrafico veniva rilevata la presenza (o l'assenza temporanea) dell'intestatario della scheda anagrafica (ISF), del suo coniuge/partner o del suo ex-coniuge/ex-partner. La menzione del coniuge/partner (anche soltanto ex) si dimostrava utile per coprire i casi di disgregazione familiare a seguito di uscita dalla famiglia dell'ISF (ad esempio per decesso o separazione) ma lasciava un margine di aleatorietà nella corretta corrispondenza tra famiglia anagrafica e famiglia di fatto. Soltanto dell'ISF, infatti, si indicavano i dati anagrafici (nome, cognome, data di nascita) mentre si lasciava al rapporto tra rispondente e rilevatore la valutazione sulla presenza del coniuge/partner.

Per cercare di rendere più oggettiva l'eleggibilità della famiglia, dal 2016, soprattutto quando le interviste vengono condotte telefonicamente, la famiglia viene considerata eleggibile se viene rilevata la presenza all'indirizzo di almeno uno dei componenti anagrafici.

Per questo motivo si chiede al rilevatore di effettuare uno screening completo della famiglia anagrafica (partendo sempre dall'ISF), rilevando la presenza in famiglia di tutti i componenti anagrafici, identificati tramite nome, cognome, sesso e data di nascita.

Per le interviste telefoniche è stata posta particolare attenzione alla verifica della corrispondenza tra indirizzo anagrafico e indirizzo effettivo. Una eventuale discordanza può infatti segnalare la presenza nelle liste anagrafiche di indirizzi non residenziali (ad esempio seconde case), che non devono essere inclusi nel campione (sovra-copertura delle liste).

Per le **famiglie delle *wave* successive** il campione è formato dalle famiglie in cui vive almeno uno dei componenti campione identificati nella prima intervista (componenti della famiglia con almeno 14 anni compiuti al 31.12 dell'anno precedente l'intervista). In sostanza, per la verifica dell'eleggibilità della famiglia viene rilevata la presenza all'indirizzo di almeno uno dei componenti-campione, identificati tramite nome, cognome, sesso e data di nascita.

Nelle *wave* successive alla prima l'indirizzo di residenza (comunque verificato) può essere differente rispetto a quello anagrafico o a quello risultante dalla precedente intervista per effetto di trasferimenti di tutta la famiglia o di qualcuno dei suoi componenti.

Fino al 2010, il **questionario individuale** che rileva informazioni su istruzione, condizioni di salute, lavoro e redditi (le informazioni socio-demografiche di base vengono raccolte per tutti i componenti della famiglia), veniva somministrato a tutti i componenti che avevano 15 anni compiuti al 31.12 dell'anno precedente l'intervista, in conformità a quanto avviene nell'indagine sulle Forze di Lavoro (salvo il fatto che per FdL l'età viene considerata al momento dell'intervista). Tuttavia, l'incremento di complessità derivante dalla contemporanea necessità di adeguarsi alle direttive europee, che fissano a 16 anni l'età minima per la rilevazione delle variabili target individuali, ha suggerito di allinearsi del tutto, dal 2011, al regolamento comunitario, escludendo i 15-enni dall'eleggibilità per il questionario individuale.

1.3 Le diverse tecniche di rilevazione utilizzate

Dal 2004 al 2010 le interviste sono state effettuate in modalità faccia a faccia con questionario cartaceo (PAPI) tramite la rete di rilevazione comunale, coinvolgendo i comuni per l'estrazione del campione di famiglie e per il reclutamento e la gestione dei rilevatori.

Gli Uffici Territoriali (UUTT) Istat sono stati coinvolti per la formazione dei rilevatori sul territorio e per eventuale supporto ai comuni nello svolgimento della rilevazione. I referenti degli UUTT sono stati precedentemente formati da esperti del servizio competente dell'indagine.

Dal 2011, le interviste sono state condotte in modalità faccia a faccia assistita da computer (CAPI). La rilevazione è stata affidata a una società esterna che ha avuto l'incarico di predisporre il software per la gestione dell'intervista (questionario elettronico) e per la produzione degli indicatori di monitoraggio della rilevazione, del reclutamento dei rilevatori, dell'organizzazione delle sessioni formative ai rilevatori (con docenza degli esperti tematici dell'Istat), della gestione del numero verde di supporto ai rispondenti, della gestione di tutto il lavoro sul campo.

I comuni sono stati coinvolti soltanto per il supporto ai rilevatori nel reperimento di informazioni sulla validità degli indirizzi e per eventuali trasferimenti di residenza.

Dal 2015, si è passati a una tecnica mista (CAPI-CATI) assistita da computer ricorrendo, oltre che alle interviste faccia a faccia, anche alle interviste telefoniche per un sottoinsieme delle famiglie campione.

L'introduzione della tecnica CATI nel 2015, motivata da ragioni emergenziali[3] ha riguardato solo un numero limitato di famiglie campione (5.269 su 17.986 totali), selezionate in base alla disponibilità di almeno un recapito telefonico (da archivio CONSODATA o da intervista precedente) e al possesso di alcune caratteristiche che le hanno fatte ritenere idonee all'utilizzo della tecnica telefonica (single o due componenti con ISF < 65 anni in prima *wave*, al massimo 2 percettori di reddito per le *wave* successive). Il questionario utilizzato è stato lo stesso della CAPI. Per l'indagine 2016 è stato pianificato un utilizzo più esteso della tecnica telefonica, che ha riguardato un sotto-campione di quasi 17.000 famiglie (per un ritorno atteso di circa 9.000 interviste) su 29.000 totali (per un ritorno complessivo atteso di circa 22.000 interviste).

Gli aspetti metodologici del passaggio alla tecnica mista e all'introduzione delle interviste telefoniche nell'indagine verranno approfonditi nel capitolo successivo.

1.4 Le misure per la prevenzione e il controllo degli errori di rilevazione

Come passo iniziale volto alla facilitazione delle operazioni sul campo, a tutte le famiglie del campione viene inviata una lettera, a firma del Presidente Istat, che le informa circa il loro coinvolgimento nella rilevazione. Nella lettera vengono espressi gli aspetti salienti dell'indagine, con particolare riferimento alla sua rilevanza ai fini delle politiche nazionali ed europee, nonché gli aspetti normativi che la regolano (obbligo di risposta, quesiti sensibili, ecc.).

È noto infatti in letteratura e confermato nei fatti che il tasso di risposta delle famiglie preavvisate in modo istituzionale circa l'intervista da effettuarsi sia sempre nettamente più elevato rispetto a quelle che, per qualsiasi motivo, non l'hanno ricevuta.

Le famiglie, inoltre, possono rivolgersi a un numero verde che può rassicurarle circa l'autenticità dell'intervista e confermare l'accredito del rilevatore. L'Istat chiede inoltre ai Comuni di farsi carico, se possibile, di inviare una loro lettera alle famiglie, a firma del Sindaco, che ulteriormente ribadisce la rilevanza dell'indagine. I Comuni stessi, infine, così come le Prefetture, vengono messi a conoscenza della rilevazione, perché le famiglie che si rivolgano loro per ottenere informazioni sull'effettivo svolgimento dell'indagine possano essere opportunamente rassicurate.

Un'attenzione particolare, poi, viene posta nel reclutamento e nell'addestramento dei rilevatori, sia che questi siano stati selezionati dai comuni (fino al 2010) oppure dalle società affidatarie che si sono avvicendate dal 2011 in avanti. Le direttive imposte dall'Istat prevedono infatti che i rilevatori posseggano determinate caratteristiche di idoneità (titolo di studio superiore, precedente esperienza) e che possano iniziare ad effettuare le interviste solo dopo avere partecipato alle sessioni formative previste prima dell'avvio della rilevazione.

Ad ogni famiglia rilevata (intervistata o meno) è inoltre associato il codice che identifica univocamente il rilevatore che l'ha avuta in carico e molti degli indicatori di qualità della rilevazione vengono dettagliati a livello di singolo rilevatore.

Nello svolgimento del lavoro sul campo, inoltre, è prevista la registrazione di un esito di rilevazione per ogni famiglia del campione. Per le famiglie che rifiutano l'intervista vengono raccolte alcune informazioni che permettono di meglio caratterizzare il rifiuto: momento del

3 Il ritardo nell'aggiudicazione della gara per l'affidamento della rilevazione 2015 ha costretto l'Istat a prorogare il contratto con la società affidataria fino al 2014, la quale ha potuto iniziare il lavoro sul campo soltanto nella prima metà di ottobre 2015. La tecnica telefonica, comunque prevista come opzione nel PSN, è stata introdotta a partire dal mese di novembre 2015 per cercare di completare la rilevazione nel più breve tempo possibile, assicurando la più adeguata copertura territoriale del campione.

rifiuto (prima di essere riusciti anche solo a parlare dell'indagine, prima di iniziare l'intervista, giunti a un determinato quesito), il motivo del rifiuto, alcune caratteristiche sommarie della persona che rifiuta.

Prima del passaggio alle tecniche di rilevazione assistite dal computer, queste informazioni, unite ad un sommario resoconto sul numero di tentativi effettuati, con o senza contatto, e sulla durata dell'intervista, fornivano soltanto gli elementi per una valutazione a posteriori della qualità della rilevazione e dell'operato dei singoli rilevatori.

Il passaggio prima al CAPI e poi al CATI, dove il flusso delle informazioni suddette avviene quasi in tempo reale, ha consentito lo sviluppo di un articolato sistema di monitoraggio della rilevazione in corso d'opera che permette, ad esempio, di intervenire con dei supplementi di formazione ai rilevatori, a distanza tramite e mail o con l'organizzazione di specifiche sessioni di *debriefing*.

La rilevazione assistita da computer, inoltre, permette di rendere più precise alcune informazioni quali la durata dell'intervista, il numero di contatti, e lo stesso esito della rilevazione.

L'utilizzo del questionario elettronico permette infatti di dare un esito anche al singolo contatto o tentativo di contatto e gli esiti di rilevazione scaturiscono automaticamente dai diversi percorsi del questionario elettronico in base a delle regole. In questo modo è possibile classificare gli esiti tenendo conto dell'intera "storia", come ad esempio le cadute con o senza contatto, con o senza appuntamento, senza tentativi di contatto (non lavorate), ecc.. La durata dell'intervista (o del singolo contatto) viene automaticamente registrata dal computer.

1.5 Sviluppo e test del questionario

La progettazione del questionario utilizzato si è basata sulle linee guida di Eurostat che definiscono le variabili target dell'indagine, sulle esperienze dell'indagine sui Bilanci di famiglia della Banca d'Italia e del Panel europeo sulle famiglie (ECHP). La versione definitiva è stata messa a punto dopo una serie di tre indagini pilota svolte tra il 2000 e il 2003 che si sono affiancate a una serie di analisi sulle fonti esistenti (Bilanci di famiglia della Banca d'Italia e Panel europeo sulle famiglie - ECHP) e di sperimentazioni ad hoc (Ceccarelli, Di Marco e Rinaldelli, 2008).

L'impianto iniziale non è stato modificato, salvo limitatissimi adeguamenti in itinere, fino al 2015.

Anche nel 2011, infatti, quando si è realizzato il passaggio alla tecnica CAPI, non è stata messa in atto una riprogettazione radicale del questionario ma soltanto la sua implementazione nel software installato sui PC dei rilevatori. E' stata comunque condotta una dettagliatissima attività di test del questionario elettronico, da parte del personale del servizio tecnico competente dell'indagine, dando luogo a tutte le correzioni necessarie, da parte della società affidataria, prima di utilizzare lo strumento per il lavoro sul campo (Freguia e Romano, 2014).

A partire dal 2016, il questionario è stato profondamente rivisto per adattarsi all'introduzione a regime della tecnica mista CAPI/CATI, come verrà più dettagliatamente mostrato nel capitolo 2. Per finire, occorre solo ricordare che per ogni edizione di indagine, a partire dal 2005, è stato necessario implementare nel questionario un modulo ad hoc di approfondimento tematico richiesto dal regolamento europeo. In questo caso, quando possibile (come ad esempio per i quesiti su "Menthal Health" del modulo 2013) le variabili target richieste dal regolamento sono state ottenute mediante quesiti standardizzati, già utilizzati in altre indagini sulle famiglie. In caso contrario è stato svolto un lavoro di adattamento delle specifiche europee al contesto nazionale e alla lingua italiana.

2. LA SPERIMENTAZIONE DELLA TECNICA CATI E LA PROGETTAZIONE DELLA RILEVAZIONE CON TECNICA MISTA CAPI/CATI[1]

In questo contributo saranno descritte la progettazione e l'attuazione della sperimentazione della tecnica CATI nell'indagine EU-SILC: sarà illustrato il contesto in cui la sperimentazione è stata avviata e le principali fasi di realizzazione, motivando le scelte prese ed evidenziando punti di forza e criticità emersi. Infine, sulla base dei risultati ottenuti, sarà presentato lo sviluppo successivo che ha portato all'implementazione della tecnica mista CAPI-CATI a regime.

2.1 Introduzione

Le politiche europee attuate negli ultimi anni hanno visto i Paesi dell'Unione Europea sempre più impegnati per il conseguimento di obiettivi di modernizzazione del Sistema statistico e di incremento della corrispondente efficienza complessiva, nonché della sua capacità di risposta alle esigenze degli utenti, in linea con il memorandum di Wiesbaden sul nuovo disegno delle statistiche sociali europee. In questo contesto, i maggiori vantaggi, anche in termini di tempestività e contenimento dei costi, si individuano nella raccolta più efficiente di dati a livello nazionale (Eupean Commission, 2011). In particolare, Euro-stat già dal 2005 raccomanda l'utilizzo di tecniche Computer Assisted (CA), poiché consentono di prevenire problemi di misura, di rendere più agevole la raccolta dei dati, nonché di identificare e correggere già durante l'intervista la maggior parte degli errori di processo "*...computer-assisted interviewing (CAPI or CATI) is definitely desirable in order to prevent measurement problems and facilitate data collection. Another advantage of computer-assisted inter-viewing is that most of the processing errors can be identified and corrected during the interview*" (Eurostat, 2008).

In Italia, in linea con tali raccomandazioni, il Programma Stat2015 e il PST 2013-2015 avevano indicato come obiettivo prioritario la progressiva modernizzazione delle statistiche sociali e delle relative rilevazioni: incrementando il ricorso a tecniche assistite dal computer, sia completando la transizione da indagini con questionario cartaceo (PAPI) a indagini con modalità CAPI, sia attivando sperimentazioni di modalità mista per limitare la numerosità della componente CAPI a favore di tecniche CA meno onerose (CAWI, CATI); sfruttando maggiormente il patrimonio informativo degli archivi amministrativi e potenziando i processi di integrazione tra dati amministrativi e dati d'indagine (Freguja, Romano, 2014).

E' in questo quadro che in Italia l'indagine EU-SILC, svoltasi dal 2004 al 2010 con tecnica PAPI e con rete di rilevazione comunale, dal 2011 ha adottato la tecnica CAPI, affidando la rilevazione ad una società esterna, e nel 2013 ha avviato la sperimentazione CATI, con l'obiettivo di testare l'applicabilità dell'intervista telefonica ad un'indagine, come è EU-SILC, caratterizzata dalla complessità dell'intervista e dalla delicatezza dei temi trattati.

La sperimentazione è stata incoraggiata, inoltre, dal fatto che 14 dei 31 Paesi che conducevano annualmente l'indagine EU-SILC utilizzavano l'intervista telefonica come tecnica

1 I paragrafi 2.1, 2.3.4, 2.3.5, 2.5 sono stati redatti da Francesca Lariccia; i paragrafi 2.2, 2.3.2, 2.3.3, 2.4.1 sono stati redatti da Doriana Frattarola, i paragrafi 2.3.1 e 2.3.6, 2.4.2 sono stati redatti da Alessandro Fratoni.

di raccolta dati: due Paesi la usavano in modo esclusivo, gli altri in combinazione con altre tecniche di rilevazione, nella maggioranza dei casi CAPI-CATI (Eurostat, 2012).

2.2 Il contesto italiano: il passaggio da CAPI a CATI

Il questionario CAPI aveva conservato l'impostazione del questionario PAPI ma aveva introdotto una gestione più efficiente delle interviste; il passaggio al questionario elettronico aveva permesso, infatti, di gestire in automatico i percorsi ed i salti tra sezioni, di "personalizzare" il testo di alcuni quesiti tramite l'utilizzo di testi mobili, nonché di effettuare controlli di compatibilità tra le informazioni rilevate, già in sede di intervista.

E' sempre con il passaggio alla tecnica CAPI che si è verificata un'altra importante novità: l'inserimento di alcuni quesiti a conferma, che nel 2011 avevano interessato solo le informazioni anagrafiche individuali, ma che negli anni successivi furono estesi ad un numero sempre maggiore di quesiti, portando miglioramenti in fase di trattamento dati, soprattutto per quanto riguarda la componente longitudinale dell'indagine (vedasi Capitolo 4.3).

La tecnica CATI mantiene tutti i vantaggi introdotti dalla CAPI e ne aggiunge altri propri della nuova tecnica. In prima istanza una riduzione di costo immediata: è evidente come il costo delle interviste telefoniche sia necessariamente più contenuto rispetto ai costi della somministrazione CAPI, che prevede spostamenti sul territorio per recarsi presso le abitazioni delle famiglie che possono essere anche notevolmente distanti tra loro. La tecnica telefonica permette, inoltre, una copertura del territorio più capillare, poiché consente di raggiungere gli intervistati anche in comuni marginali e in situazioni che rendono difficile recarsi presso l'abitazione della famiglia campione (condizioni climatiche avverse, calamità naturali, ecc.). Per ragioni simili, la tecnica CATI porta con sé anche una notevole riduzione dei tempi di realizzazione delle interviste totali, potendo effettuare, in linea teorica, un maggior numero di interviste al giorno telefonicamente, non dovendo fisicamente raggiungere le famiglie; inoltre il contatto telefonico prevede la possibilità di superare le reticenze mostrate da alcune famiglie relative all'"accogliere fisicamente" in casa propria un rilevatore. L'intervista telefonica comporta anche una riduzione del *response burden*, dato che risulta, sotto diversi profili (psicologico, logistico, ecc.), meno impegnativa per il rispondente.

Inoltre, l'intervista telefonica permette di effettuare un più attento, efficiente ed efficace monitoraggio dell'attività dei rilevatori: essi sono solitamente concentrati in un'unica sede e possono essere seguiti direttamente e in tempo reale, mentre svolgono le interviste, rendendo possibili interventi migliorativi sulle modalità di conduzione, in linea con contenuti tematici e regolamento europeo, nonché interventi correttivi su eventuali comportamenti non conformi alle metodologie previste dal disegno di indagine.

Tuttavia, le interviste telefoniche presentano anche dei limiti e degli svantaggi. Innanzitutto possono essere intervistate in questa modalità solo le famiglie per cui si ha a disposizione il recapito telefonico. Un secondo aspetto da considerare riguarda il venir meno di un rapporto "diretto" tra rilevatore e intervistato, e questo riduce la possibilità per il primo di cogliere e quindi risolvere eventuali incertezze dell'intervistato. Il contatto telefonico, in definitiva, rende sicuramente più complicato un supporto adeguato alle famiglie durante l'intervista, svantaggio che solo una attenta e adeguata formazione preliminare dei rilevatori può contenere.

Così come accadeva con la tecnica CAPI, anche con la CATI la fase di inserimento dati è contestuale all'intervista e le risposte vengono inserite direttamente al computer. La gestione delle chiamate, così come degli appuntamenti, avviene in modo automatico e

organizzato via *software*. Tutto ciò garantisce maggiore precisione nella somministrazione del questionario pur preservando l'importanza del ruolo dell'intervistatore, che dovrà essere adeguatamente preparato per riuscire a instaurare un rapporto "di fiducia" con il rispondente. Quest'ultimo, reso consapevole del suo ruolo, dovrà sentirsi stimolato a partecipare all'indagine e, nello stesso tempo, sentirsi supportato dal rilevatore nello svolgimento dell'intervista. Anche in questo senso la fase di formazione dei rilevatori assume una fondamentale importanza.

Facendo le prime riflessioni circa la possibilità di somministrare il questionario EU-SILC in modalità CATI sono emersi da subito altri punti critici.

In linea generale, la durata di un'intervista telefonica dovrebbe essere il più possibile contenuta; è infatti molto difficile mantenere alta l'attenzione dei rispondenti su argomenti complessi e delicati, quali possono essere considerati quelli dell'indagine EU-SILC, stando al telefono. Considerando che la durata media delle interviste in modalità CAPI nell'edizione 2011 dell'indagine era risultata di ben 46 minuti, si avvertiva il rischio che la tecnica CATI avrebbe potuto produrre un considerevole aumento dei rifiuti da parte delle famiglie, anche dopo l'avvio dell'intervista, che normalmente è il momento in cui si verifica la maggior parte dei rifiuti.

Anche riuscendo a condurre a termine l'intervista, inoltre, esisteva il timore di una maggiore imprecisione nella rilevazione, per un'eventuale maggiore sbrigatività dei rilevatori o dei rispondenti, i quali avrebbero potuto facilmente rinunciare a consultare documenti utili a rispondere in modo preciso ad alcuni quesiti sui loro redditi o sulle spese della famiglia.

Si temeva infine un maggiore ricorso alla modalità *proxy*, cioè a concludere l'intervista per interposta persona, a causa della maggiore difficoltà a coinvolgere direttamente gli interessati, dei quali non si può sapere, al telefono, se al momento dell'intervista sono presenti o meno al domicilio.

La sperimentazione CATI è nata proprio con l'obiettivo di verificare la fattibilità dell'introduzione di questa tecnica per l'indagine EU-SILC nel contesto italiano, mettendo in opera tutti gli accorgimenti utili a contenere le criticità evidenziate e metterne in luce eventuali altre non previste.

2.3 La sperimentazione CATI

2.3.1 Organizzazione della sperimentazione e selezione del campione

Al fine di sperimentare l'adozione della tecnica CATI nell'indagine EU-SILC, è stata affidata la conduzione, la gestione ed il monitoraggio di una rilevazione pilota a una Società esterna, la quale ha curato l'implementazione del questionario elettronico, ha messo a disposizione le risorse umane (trenta rilevatori e due supervisori di sala) e un Numero Verde, cioè un servizio telefonico di supporto alle famiglie attivo per tutta la durata della sperimentazione.

I rilevatori e i supervisori di sala, nonché gli addetti al Numero Verde, sono stati opportunamente formati sui contenuti dell'indagine e sulla metodologia di conduzione delle interviste da personale Istat.

L'obiettivo della rilevazione era la realizzazione di 1.500 interviste telefoniche complete a partire da un campione di circa 6000 famiglie, composto:

- per metà da "nuove" famiglie, estratte dalle Liste Anagrafiche Comunali(LAC);

- per il resto da famiglie che avevano già partecipato all'indagine EU-SILC e che avevano effettuato la quarta e ultima intervista nel 2011 o nel 2012; per queste famiglie si è trattato quindi della "quinta" occasione di indagine.

All'interno di ciascun gruppo, in modo ragionato, le famiglie sono state ulteriormente selezionate, escludendo quelle potenzialmente portatrici di una eccessiva complessità di intervista. Per quanto riguarda il primo gruppo, sono state selezionate le famiglie formate da un componente, oppure da due componenti e intestatario della scheda di età uguale o superiore a 65 anni. Per quanto riguarda il secondo gruppo sono state selezionate le famiglie composte da un numero massimo di due percettori di reddito. Queste condizioni di selezione dovevano essere valide all'estrazione del campione: ovviamente in diverse occasioni si è verificata una discrepanza con la composizione effettiva della famiglia al momento dell'intervista, che comunque andava intervistata nella sua interezza. Trattandosi di una indagine pilota, la partecipazione da parte delle famiglie non è stata obbligatoria.

Il campione è stato organizzato in quartine, con una famiglia base e tre sostitutive, da utilizzare secondo specifiche regole nei casi di caduta della famiglia base. La lettera informativa alle famiglie è stata inviata solo alla famiglia base della quartina e alla prima sostituta.

2.3.2 Riprogettazione del questionario

La sperimentazione CATI ha preso le fila da una riprogettazione del questionario, per renderlo più rispondente alla nuova tecnica di somministrazione: i quesiti dovevano essere brevi, chiari ed efficaci, ancor più di quanto già non lo fossero per la tecnica CAPI, dove l'intervista avveniva alla presenza del rilevatore.

In primo luogo, grazie all'aiuto dei colleghi esperti del settore, sono state analizzate le schede di *de-briefing* compilate dai rilevatori CAPI che avevano lavorato nelle indagini EU-SILC più recenti, allo scopo di risolvere eventuali problematiche evidenziate e accogliere i suggerimenti proposti.

È stata poi effettuata un'attenta verifica del *wording* dei quesiti, per migliorarne la comprensibilità e l'efficacia, soprattutto in relazione ad una loro lettura telefonica, salvaguardando i contenuti informativi imposti dal regolamento europeo: per esempio sono stati eliminati molti incisi che sono invece rimasti presenti negli *help online* e consultabili solo in caso di necessità.

Inoltre, per rendere più fluida e scorrevole l'intervista, sono state riviste le sequenze di intere sezioni del questionario (ad esempio la sezione sulle caratteristiche dell'occupazione) per migliorare la coerenza dei percorsi e dei quesiti alla luce delle risposte fornite ai quesiti precedenti.

Un'attenzione particolare è stata inoltre dedicata a verificare quali quesiti potessero essere eliminati, in quanto volti a cogliere informazioni desumibili da archivi amministrativi.

A tal fine è stato innanzitutto necessario operare una distinzione tra gli individui del campione già presenti nella famiglia anagrafica e quelli presenti solo nella famiglia di fatto. Soltanto per i primi[2], infatti, risulta praticabile il loro collegamento alle fonti amministrative, tramite il codice fiscale presente nelle LAC (Liste Anagrafiche Comunali) dalle quali

2 Per questioni operative i codici fiscali degli individui del campione, necessari per il collegamento con i dati amministrativi, devono essere forniti al settore dell'Istat competente per la raccolta delle fonti amministrative, con un certo lasso di tempo in anticipo rispetto al momento in cui tali fonti diventano disponibili e utilizzabili nell'ambito del processo produttivo di EU-SILC, a causa delle necessarie operazioni di estrapolazione delle informazioni dagli archivi amministrativi stessi. Ciò implica che la fornitura dei codici fiscali non possa includere anche gli individui di fatto aggiuntivi, per i quali la soluzione alternativa è stata la previsione della somministrazione di un'intervista completa.

il campione è stato estratto. Per i componenti di fatto della famiglia, si è invece deciso di prevedere un percorso di intervista completo.

È stato così stabilito di non somministrare agli individui anagrafici la sezione del questionario individuale dedicata alle pensioni, ad eccezione di quelle erogate da un ente previdenziale estero, e sono stati anche eliminati i quesiti relativi a altre tipologie di trasferimenti sociali (CIG, disoccupazione, assegni al nucleo familiare), anche essi desumibili da archivi amministrativi. Tutte queste parti del questionario sono invece state mantenute per i componenti di fatto delle famiglie.

Per tutti gli individui, anagrafici o di fatto, è stata comunque mantenuta la rilevazione di un prospetto delle diverse tipologie di reddito percepite, precedente la rilevazione delle informazioni di dettaglio e degli importi percepiti, utile in fase di intervista per la gestione dei percorsi e in fase di trattamento per la riconciliazione con i dati amministrativi e/o per l'imputazione dei dati di dettaglio mancanti.

Infine, la sperimentazione CATI è stata l'occasione per testare una doppia modalità di rilancio nel caso di risposte evasive da parte dell'intervistato ("Non sa", "Non risponde") ad importanti quesiti reddituali.

Già prima della sperimentazione, nel questionario veniva proposto, solo per un piccolo gruppo di variabili cruciali, un secondo quesito volto a cogliere almeno un importo in classi, qualora l'intervistato non rispondesse al quesito relativo all'importo puntuale. Nell'ottica di rendere il questionario più snello, nel corso della progettazione CATI, sono state fatte riflessioni sulla possibilità di eliminare questo secondo quesito, o di mantenerlo in quanto effettivamente efficace per recuperare una parte di informazione che altrimenti sarebbe andata persa.

Inoltre, con l'avvento del questionario elettronico, per ciascuna variabile quantitativa, è stata prevista una modalità "standard" di rilancio: nel momento in cui l'intervistato non risponde, per il rilevatore è visualizzato a video il seguente messaggio di *warning*: "Poiché l'informazione è importante, può darmi almeno un'indicazione approssimativa?" che lo guida nel tentativo di avere comunque una risposta dall'intervistato. E' stato deciso, dunque, di testare se quest'ultima potesse essere la modalità di rilancio migliore per tutti i quesiti quantitativi, ivi compresi quelli del primo gruppo di variabili fondamentali, cui sono stati aggiunti anche altri quesiti del questionario individuale (ad esempio relativi alle perdite dei lavoratori autonomi). Alle famiglie è stata associata una variabile *flag* che individuasse il tipo di rilancio presente durante il corso di tutta l'intervista: il primo tipo di rilancio consisteva nel sollecitare l'intervistato tramite il *warning* proposto al rilevatore; il secondo tipo, invece, non prevedeva questo sollecito verbale, bensì la somministrazione di un secondo quesito per cogliere l'importo in classi reddituali. I due diversi rilanci sono stati gestiti in automatico dal questionario elettronico.

L'obiettivo è stato dunque stabilire a posteriori quale modalità di rilancio tra le due fosse più efficace e rispondente a favorire la rilevazione di alcune tipologie di reddito anche raccogliendo importi approssimati.

Il questionario CATI, riprogettato come descritto sopra, è stato implementato dalla Società incaricata in formato elettronico interattivo, sulla base di specifiche stabilite dall'Istat relativamente a:

- il piano di compatibilità, che prevedeva controlli di *range* per alcune variabili, regole di coerenza tra le notizie anagrafiche dei componenti della famiglia e regole di coerenza tra le risposte date dall'intervistato durante la rilevazione, nonché un sistema di *warning* che venivano visualizzati nel momento di violazione di tali regole per essere lette dal rilevatore all'intervistato e eventualmente correggere l'incoerenza riscontrata;

- la navigazione agile del questionario, che consentiva la possibilità per l'intervistatore di tornare indietro al quesito che ha originato una possibile incoerenza ed eventualmente rettificarne le risposte;
- l'"*Help in linea*", con la possibilità di visualizzare immediatamente nel corso dell'intervista le schermate di *help* per i quesiti che lo prevedevano;
- la codifica assistita di determinate variabili tramite l'uso di menù a tendina o di appositi motori di ricerca per la classificazione della professione e del settore di attività economica dei lavoratori;
- la gestione delle domande a conferma, utilizzando e riconciliando le informazioni acquisite nelle precedenti occasioni di indagine;
- l'impostazione grafica - ad esempio l'uso di colori o di diverse dimensioni del carattere nelle diverse parti del questionario e la visualizzazione di più quesiti nell'ambito di una stessa schermata, la possibilità di specificare indicazioni per il rilevatore, utilizzando sempre lo stesso carattere (corsivo) e la stessa posizione (subito sotto il quesito e preceduto dall'indicazione "Per il rilevatore:");
- la gestione dell'agenda dell'intervistatore per gli appuntamenti con le famiglie (selezione automatica casuale dei nominativi ed assegnazione automatica agli intervistatori; gestione automatica degli appuntamenti, dei tentativi di contatto, dei richiami telefonici, degli esiti di contatto);
- la capacità di memorizzare variabili proprie della rilevazione e di sistema, in maniera automatica e trasparente (ad esempio, esiti, minuti, ora e data dei tentativi di contatto, minuti, ora e data di inizio e fine intervista);
- l'elaborazione degli indicatori necessari alla gestione e al monitoraggio dell'indagine;
- la produzione dei file di dati e i meccanismi di protezione dei dati in grado di gestire i profili di accesso alle utenze autorizzate.

Il sistema applicativo utilizzato ha consentito, inoltre, di apportare modifiche al questionario elettronico anche durante la rilevazione sulla base di eventuali problemi emersi nei contenuti, nei percorsi di intervista o nell'impostazione grafica.

2.3.3 Test del questionario

Avendo apportato importanti modifiche al questionario, particolare attenzione è stata dedicata alla fase di test del questionario stesso.

La società selezionata per sviluppare il questionario elettronico CATI è stata di fatto la stessa società che aveva curato, fino a quel momento, l'implementazione del questionario CAPI. Per questo motivo l'implementazione di tutte le modifiche richieste è avvenuta su uno "scheletro" già esistente, il questionario EU-SILC 2011 CAPI. Ciò ha consentito di ridurre i tempi di implementazione del questionario CATI, ma non ha evitato la necessità di verificare con attenzione la sua completa funzionalità, oltre che la corretta implementazione delle modifiche richieste, ponendo particolare attenzione a che queste stesse modifiche non avessero introdotto anomalie di funzionamento anche in parti del questionario non interessate da cambiamenti.

Sono stati oggetto di un controllo particolarmente approfondito:

- la correttezza del *wording*;
- l'adeguatezza del *layout*;
- la congruenza delle regole di incompatibilità dopo la revisione del questionario;
- la coerenza dei percorsi con quanto richiesto in fase di progettazione;
- la corretta implementazione dell'"*Help in linea*", opportunamente adattato alla nuova tecnica;
- le funzionalità dell'agenda dell'intervistatore.

Questa intensa fase si è svolta nell'arco di due mesi (dicembre 2013 - gennaio 2014): la società esterna ha messo a disposizione alcune utenze attraverso le quali è stato possibile simulare delle interviste complete.

2.3.4 Formazione dei rilevatori

Il corso di formazione dei rilevatori reclutati dalla società incaricata si è svolto in tre giornate, dal 30 gennaio al 3 febbraio 2014. Sono stati formati circa trenta rilevatori, effettivamente impegnati nelle interviste; hanno partecipato, inoltre, due addetti al numero verde e due supervisori di sala. La frequenza all'intero corso è stata per tutti obbligatoria.

Il corso è stato anche l'occasione per testare il processo stesso di formazione dei rilevatori con tecnica CATI e avere, durante la rilevazione, indicazioni di ritorno sulla sua efficacia, tramite il monitoraggio di sala e alcuni *de-briefing* opportunamente organizzati durante lo svolgimento della rilevazione. Inoltre, il numero contenuto di partecipanti ha consentito di sperimentare una formula innovativa di formazione, più breve nella durata, e più snella e operativa nei contenuti e nelle modalità di svolgimento rispetto a quella abitualmente condotta per formare i rilevatori CAPI di EU-SILC. Tale formula innovativa è stata caratterizzata da moduli più brevi in cui teoria ed esercitazione sono stati integrati in un'unica sessione, anziché in sessioni distinte. La nuova strategia di formazione dei rilevatori è stata poi utilizzata anche per la rilevazione corrente di EU-SILC a partire dall'indagine del 2014, seppure ancora condotta con tecnica CAPI.

Il corso di formazione è stato realizzato dal personale ISTAT competente dell'indagine alla presenza del responsabile di progetto della Società esterna e ha avuto lo scopo di formare gli intervistatori sulle finalità e i modi di operare dell'Istituto, gli obiettivi dell'indagine, la struttura del questionario, il significato delle domande, la corretta esecuzione dell'intervista, i comportamenti da tenere durante l'intervista e con l'intervistato, le strategie da attivare per motivare le famiglie a collaborare all'indagine e per contenere i rifiuti. Le parti tecniche riguardanti il funzionamento del questionario elettronico e la gestione dell'agenda sono state condotte dalla Società incaricata, alla presenza del personale ISTAT competente dell'indagine. Per gli operatori del numero verde la partecipazione al corso di formazione ha avuto come obiettivo la conoscenza dell'organizzazione e dei contenuti della rilevazione in modo da avere elementi utili: a rassicurare le famiglie su eventuali dubbi o interrogativi inerenti all'indagine, a motivare le famiglie restie a collaborare, a gestire la scheda informatizzata per la registrazione dei dati delle famiglie che hanno chiamato il numero verde.

I diversi moduli formativi sono stati impostati su un piano comunicativo facilmente accessibile ai rilevatori, finalizzato a trasmettere loro la padronanza del questionario e la capacità di gestire agevolmente i singoli quesiti, nonché le modalità più adeguate per creare un clima di fiducia e di collaborazione con le famiglie. Il numero contenuto dei partecipanti, la possibilità di gestire nella stessa aula sia i momenti esercitativi sia quelli di approfondimento teorico, ha permesso di svolgere una formazione fluida e coinvolgente, in grado di favorire l'interazione con i partecipanti e di suscitare interesse per l'indagine.

In concreto, dopo aver presentato in generale finalità e temi dell'indagine EU-SILC e aver illustrato l'organizzazione della rilevazione e la struttura del questionario, il corso si è articolato conducendo i rilevatori lungo il percorso di un'intervista completa: dalla scheda contatti fino al completamento dei questionari individuali. Ogni modulo dedicato al questionario (scheda contatti, scheda generale, questionario familiare, sezioni del questionario individuale), dopo una breve introduzione dedicata a illustrarne i principali contenuti infor-

mativi, è stato svolto alternando la conduzione dell'esercitazione, quesito per quesito, delle sezioni coinvolte, a momenti di approfondimento teorico sulle informazioni più rilevanti o di più difficile rilevazione

Un modulo a sé stante prevedeva un focus sul lavoro, dedicato alla classificazione delle professioni e delle attività economiche, in modo da presentare la logica delle due classificazioni e fornire gli elementi per una corretta codifica. Anche in questo caso il modulo ha integrato in un'unica sessione contenuti informativi ed esercitazioni.

A completamento del percorso formativo, sono stati proposti alcuni moduli mirati a fornire ai rilevatori strumenti, strategie, tecniche e norme di comportamento per condurre l'intervista telefonica in modo appropriato, corretto ed efficace. Tali moduli hanno riguardato: la somministrazione nell'intervista CATI, in particolare come si leggono i quesiti, cosa fare e non fare durante l'intervista, come codificare le risposte; la comunicazione efficace al telefono, focalizzandosi sull'analisi del processo comunicativo, su come governarlo e come conquistare la fiducia delle famiglie e gestirne le domande e i rifiuti; il *role playing*, ossia una simulazione del contatto con gli intervistati per migliorare la comunicazione.

Nell'ultimo modulo, come esercitazione conclusiva, è stata simulata in aula la somministrazione di un'intera intervista. A seguire, ogni rilevatore si è potuto esercitare autonomamente con alcune interviste di prova in modo da acquisire padronanza tecnico-contenutistica nella somministrazione del questionario. Questo esercizio ha avuto lo scopo di migliorare la capacità persuasiva dell'intervistatore, di accrescerne l'abilità nello stabilire il rapporto di fiducia con l'intervistato e di far acquisire padronanza nell'utilizzo e gestione del questionario elettronico.

A conclusione del percorso formativo, i rilevatori sono stati sottoposti ad un test di verifica, compilato singolarmente e discusso collettivamente in aula, al fine di consolidare i principali contenuti della formazione e verificarne l'apprendimento.

Infine, per valutare l'efficacia della formazione e migliorare le edizioni future, ai rilevatori è stato chiesto di esprimere una valutazione dell'intervento formativo compilando una scheda al termine del corso.

Prima della formazione è stato fornito ai rilevatori del materiale cartaceo utile a sfruttare al meglio il percorso formativo, completare la loro preparazione e condurre le interviste (cronogramma della formazione, questionario di indagine, lettera informativa alle famiglie, guida per l'intervistatore, test di autoverifica dell'apprendimento, test di valutazione del corso), oltre agli strumenti di supporto informatico ("*Help in linea*", regole di controllo, navigatori delle classificazioni, ecc.). Ciascun intervistatore ha avuto a disposizione un PC portatile dove era caricato il questionario elettronico con cui condurre le esercitazioni relative ai diversi moduli e le interviste di prova.

2.3.5 Rilevazione, sistema di indicatori di monitoraggio, monitoraggio di sala

Rilevazione

La rilevazione CATI si è svolta nell'arco di circa un mese, dal 4 febbraio al 6 marzo 2014. Le interviste sono state effettuate tutti i giorni della settimana, tranne la domenica, all'interno di tre turni di lavoro giornalieri, in modo da coprire l'intera giornata dalle 10 di mattina alle 21, ad eccezione del sabato in cui le attività di rilevazione sono state concluse alle 17.30. Per ciascun turno di intervista sono stati coinvolti al massimo 25 intervistatori, collocati in due sale contigue, con una progressiva riduzione del numero di rilevatori coinvolti al ridursi della numerosità delle famiglie ancora da intervistare.

La rilevazione ha coinvolto, oltre agli intervistatori, un responsabile di *field*, alcuni responsabili di sala, e alcuni addetti al numero verde. Più nel dettaglio, il responsabile di *field* è una figura esperta nelle attività di reclutamento, coordinamento e supervisione dei responsabili di sala e degli intervistatori, nonché nella gestione dell'indagine, con compiti di coordinamento e di supervisione di tutte le attività operative legate all'indagine telefonica, per ottimizzare il lavoro degli intervistatori e minimizzare gli errori "non campionari" prodotti nella fase delle interviste: l'organizzazione e il controllo dei turni di lavoro; il monitoraggio dell'andamento dell'indagine in termini quantitativi e qualitativi, in modo da poter intervenire in corso d'opera a sanare e risolvere eventuali problemi. I responsabili di sala (supervisori), invece, si sono occupati del supporto in sala agli intervistatori, in merito agli aspetti sia tecnici sia contenutistici del questionario elettronico. Gli operatori addetti al numero verde, infine, hanno garantito la copertura del servizio di risposta alle famiglie a partire dalla settimana precedente l'inizio della rilevazione e fino al completamento dei lavori.

Le interviste sono state effettuate al recapito telefonico disponibile all'avvio dei lavori o a qualsiasi altro numero, fisso o di cellulare, che l'intervistato abbia eventualmente indicato a seguito del primo contatto telefonico. Nei casi in cui non sia stato possibile completare l'intervista con un solo contatto telefonico, si è proceduto con ulteriori contatti. L'intervista alla famiglia è stata considerata completa solo dopo aver acquisito, per tutti i componenti eleggibili, i dati relativi a tutte le sezioni previste nel questionario.

Il database è stato conservato su supporto magnetico centralizzato e soggetto a strategie di *backup*. Ogni giorno il sistema provvedeva ad aggiornare automaticamente i dati residenti nel sistema CATI e a generare di conseguenza file riguardanti i tentativi di contatto e i contatti con esito definitivo, e report di elaborazioni sulle famiglie, sull'attività degli intervistatori, sugli esiti provvisori e definitivi classificati secondo alcune variabili di interesse. Per tutta la durata della rilevazione sono stati svolti controlli di qualità al fine di verificare la rispondenza dei requisiti della fornitura e di correggere, in corso d'opera, eventuali criticità in modo da raggiungere gli obiettivi prefissati. Il monitoraggio della qualità è stato realizzato dal personale Istat competente dell'indagine sia attraverso l'analisi di indicatori specifici prodotti giornalmente dalla Società incaricata, sia attraverso la supervisione in sala delle interviste per controllare il regolare svolgimento delle operazioni di raccolta dei dati e per fornire supporto ai supervisori nella risoluzione di eventuali problemi non previsti o particolarmente complessi, interagendo anche con i rilevatori.

Sistema indicatori di monitoraggio

Grazie alla fornitura, pressoché contestuale all'intervista, di informazioni che si aggiungono a quelle raccolte attraverso i questionari, il monitoraggio consente di avere un quadro aggiornato sull'andamento della rilevazione e migliora in tempo reale la qualità dell'indagine. L'elaborazione dei dati di monitoraggio, disaggregati anche per territorio e per rilevatore, infatti, permette di intervenire tempestivamente e in modo mirato sulle eventuali criticità riscontrate. Per la sperimentazione CATI, il *software* ha consentito la produzione e la trasmissione giornaliera di indicatori di qualità dell'indagine per tutto il periodo di svolgimento della rilevazione.

In particolare, il sistema di monitoraggio prevedeva i seguenti indicatori: stato delle famiglie contattate (in termini di esito definitivo), principali tassi di contatto delle famiglie (per intervistatore, per tipologia familiare e ricezione della lettera, per territorio, e cumulata totale). Tutti gli indicatori sono stati declinati anche per giorno e per settimana di calendario. Le schede di monitoraggio prevedevano anche: indicatori giornalieri per rilevatore su

durata dell'intervista, tentativi di contatto e interviste *proxy*, declinati per tipologia familiare e numero di componenti. Dato che la sperimentazione CATI è stata rivolta a famiglie alla prima e alla quinta occasione di indagine, è stato necessario calcolare gli indicatori anche per *wave* di appartenenza.

Monitoraggio di sala

Il monitoraggio di sala è stato finalizzato principalmente a cogliere eventuali criticità della somministrazione, i passaggi di maggiore difficoltà del questionario nonché all'individuazione delle resistenze più frequenti espresse dalle famiglie intervistate.

L'attività è stata organizzata in turni, in modo da garantire una presenza nel maggior numero possibile di giorni. Per valutare aspetti differenti della conduzione delle interviste e del comportamento dei rilevatori, durante ogni turno è stato possibile essere presenti in sala come osservatore non partecipante, avendo anche la possibilità di ascoltare in cuffia le telefonate. Alla fine di ogni turno, le osservazioni maggiormente rilevanti sono state riportate in apposite schede di monitoraggio. In particolare, la "scheda questionario", con griglia articolata per sezioni e per turno di monitoraggio, permetteva di annotare, per singoli quesiti o per batterie di domande, difficoltà riscontrate sui concetti, sulla formulazione dei quesiti, sulla terminologia adottata; inoltre, in uno spazio dedicato, era possibile inserire delle annotazioni di tipo generale sul questionario: particolari problemi tecnici del *software*, eventuale presenza di errori/problemi sistematici nel percorso del questionario, tipologie di famiglie (ad esempio per numero di componenti e per *wave*) che creavano maggiori problemi, frequenti interruzioni di interviste e/o rifiuti. La "scheda intervistatore", invece, con griglia articolata per rilevatore e per turno di monitoraggio, consentiva di segnalare osservazioni sui singoli rilevatori. Infine, la "scheda risposte", con griglia articolata per sezioni di questionario, permetteva di scrivere eventuali domande sorte nel corso del turno e le relative risposte fornite, in modo da condividere chiarimenti e approfondimenti forniti con il resto del personale coinvolto (rilevatori, supervisori, personale Istat competente per l'indagine).

Data la natura sperimentale della rilevazione, si è ritenuto importante organizzare dei *de-briefing* con gli intervistatori in modo da sfruttare la loro esperienza *in itinere*, riportando le difficoltà principali incontrate nel lavoro sul campo.

Il primo *de-briefing* con i rilevatori si è svolto dopo la prima settimana di lavoro sul campo ed è stato condotto dal personale ISTAT competente dell'indagine alla presenza del responsabile di progetto della Società incaricata e dei responsabili di sala in modo che le varie figure coinvolte fossero al corrente delle diverse problematiche affrontate. Durante il *de-briefing*, svoltosi negli stessi locali dove si svolgeva la rilevazione, sono stati ripresi alcuni dubbi nell'interpretazione di quesiti o alcune richieste di chiarimento, emersi durante i primi giorni di attività di sala. In particolare, a seguito di segnalazioni intervenute da parte dei rilevatori e/o del personale ISTAT che aveva svolto l'attività di monitoraggio, si è ritenuto utile condividere con tutto il gruppo degli intervistatori i chiarimenti e le raccomandazioni, fornite di volta in volta ai singoli operatori telefonici. La scheda di *de-briefing* è stata così strutturata: quesito del questionario, domanda/richiesta di chiarimento, risposta/spiegazione/esempi, riferimento alla guida per l'intervistatore.

2.3.6 I risultati della sperimentazione

L'analisi dei risultati ottenuti è stata effettuata confrontando i dati della sperimentazione CATI con quelli dell'indagine ordinaria precedente, EU-SILC 2013, condotta interamente con tecnica CAPI. Per garantire la piena confrontabilità, per la sperimentazione CATI sono state prese in considerazione solo le famiglie base e le prime sostitute, a cui è stata inviata la lettera informativa, mentre per la CAPI 2013 sono state considerate solo le famiglie con le stesse caratteristiche di quelle incluse nel campione CATI, così come descritto nel paragrafo 3.1.

Sono inoltre state escluse dall'analisi tutte le famiglie per le quali non era stato conseguito un esito definitivo; quindi non vengono incluse nel totale:

- le famiglie che durante la sperimentazione CATI sono state contattate (anche più volte, in quanto, in corso di rilevazione, viste le difficoltà riscontrate, sono state concesse deroghe al numero di contatti possibili e ai criteri di sostituzione) e per le quali sono stati registrati solo contatti con telefono libero, occupato o irraggiungibile;
- le famiglie che non sono state mai contattate dal rilevatore, sia CAPI che CATI.

Dall'analisi degli indicatori di esito dell'intervista si evidenziano alcune criticità per la tecnica CATI (Tavola 2.1). Appare evidente la forte differenza nel tasso di completezza, inferiore sia per le famiglie di prima *wave* (48,7% nella sperimentazione CATI e 64,9% nella CAPI 2013), che per le famiglie di *wave* successive (56,4% e 85,6% rispettivamente). Tale differenza è accompagnata da un forte aumento, per la tecnica CATI, di rifiuti e interruzioni definitive ad intervista iniziata, anche in questo caso sia per la prima *wave*, che per le *wave* successive.

Occorre ricordare che, nella sperimentazione CATI, con "*wave* successiva alla prima" si intendono le famiglie del campione selezionate da quelle che avevano già concluso i quattro anni di intervista previsti dall'indagine EU-SILC. Ciò ha comportato, in molti casi, un ostacolo alla collaborazione di queste famiglie, le quali alla fine hanno rifiutato di partecipare alla sperimentazione.

Tavola 2.1 - Famiglie per esito dell'intervista, tecnica e *wave* (composizione percentuale per esito e valori assoluti)

ESITO	EU-SILC 2013 (CAPI)			SPERIM. 2014 (CATI)		
	Wave 1	Wave 2+	TOTALE	Wave 1	Wave 2+	TOTALE
Interviste complete	64,9	85,6	80,0	48,7	56,4	53,2
Interviste incomplete:						
- rifiuto	*22,6*	*11,3*	*14,3*	*36,9*	*27,5*	*31,4*
- nome/indirizzo/telefono sconosciuto	*4,1*	*1,0*	*1,9*	*12,0*	*12,4*	*12,2*
- altri motivi di caduta	*8,3*	*2,2*	*3,8*	*2,4*	*3,7*	*3,1*
Totale famiglie	**4.385**	**12.029**	**16.414**	**702**	**985**	**1.687**

Si registra inoltre, con la tecnica CATI, un leggero aumento delle interviste *proxy* imputabile alla famiglie di prima *wave*, in particolare con almeno tre componenti (Tavola 2.2):

Tavola 2.2 - Interviste *proxy* per numero di componenti della famiglia, tecnica e *wave* (valori per 100 interviste individuali)

NUMERO COMPONENTI	EU-SILC 2013 (CAPI)			SPERIM. 2014 (CATI)		
	Wave 1	Wave 2+	TOTALE	Wave 1	Wave 2+	TOTALE
2	26,2	29,8	29,1	32,3	32,8	32,6
3 o più	43,0	48,5	48,4	56,7	43,1	44,9
Totale	**27,2**	**37,3**	**36,0**	**35,0**	**37,3**	**36 ,4**

Si riscontra, infine, una diminuzione della durata media dell'intervista per le famiglie di uno o due componenti, a prescindere dalla *wave* di appartenenza; per le famiglie più numerose, invece, la tecnica telefonica non riduce i tempi dell'intervista (Tavola 2.3).

Tavola 2.3 - Durata media delle interviste (in minuti) per numero di componenti della famiglia, tecnica e *wave*

NUMERO COMPONENTI	EU-SILC 2013 (CAPI)			SPERIM. 2014 (CATI)		
	Wave 1	Wave 2+	TOTALE	Wave 1	Wave 2+	TOTALE
1	36	34	35	31	29	30
2	47	46	46	40	42	41
3 o più	62	54	54	67	54	55
Totale	**41**	**43**	**43**	**40**	**42**	**41**

2.4 Rilevazione con tecnica mista CAPI/CATI

2.4.1 La progettazione della tecnica mista CAPI/CATI

La sperimentazione CATI è nata dall'esigenza di ridurre i costi senza limitare la qualità delle statistiche sulle condizioni di vita. Gli indicatori di qualità della sperimentazione e il feedback ricevuto dai rilevatori hanno evidenziato che l'intervista telefonica, non senza alcune criticità, ha funzionato.

In ogni caso questa sperimentazione ha condotto a valutare la tecnica CATI non del tutto sostituibile alla CAPI a causa della non completezza dei recapiti telefonici disponibili, dei problemi di contatto con le famiglie di prima *wave* e dei tassi di risposta inferiori. Inoltre la lunghezza del questionario rende gravoso l'utilizzo dell'intervista telefonica per le famiglie con più percettori di reddito. D'altra parte la tecnica CAPI riesce con difficoltà a coprire le aree del Paese meno accessibili, comporta un lavoro sul campo più lungo e consente una minore possibilità di monitoraggio dei rilevatori.

Alla luce di queste considerazioni, a partire dall'edizione 2016 l'indagine è stata progettata per essere realizzata con tecnica mista CAPI/CATI, con l'obiettivo di sfruttare al massimo i rispettivi vantaggi offerti dalle due tecniche.

L'indagine prevede che la rilevazione sia condotta in modo simultaneo con le due tecniche. Il campione è stato distribuito per tecnica sulla base della disponibilità dei numeri di telefono. Più in particolare, è stato adottato il seguente criterio di base:

- intervistare in modalità CATI tutte le famiglie per le quali si disponeva di almeno un recapito telefonico, indipendentemente dalle dimensioni della famiglia (la tecnica telefonica quindi non è stata utilizzata per un sottocampione selezionato di famiglie, come nel caso della sperimentazione);
- utilizzare la modalità CAPI per tutte le famiglie per le quali non si disponeva di alcun recapito telefonico e per tutte le famiglie in cui l'intestatario della scheda era di nazionalità straniera, indipendentemente dalla disponibilità del numero di telefono.

Non è stata lasciata alle famiglie la scelta tra le due tecniche di rilevazione ma è stato comunque previsto di effettuare spostamenti di nominativi tra le due tecniche durante la rilevazione, secondo le necessità emerse dal lavoro sul campo[3].

3 In specifici casi in cui la famiglia ha espresso delle insormontabili resistenze ad una intervista telefonica, è stato

Dato che la Società incaricata è risultata essere diversa da quella che aveva condotto la rilevazione fino all'anno precedente, nonché la sperimentazione CATI, il questionario elettronico ha dovuto essere sviluppato ex-novo, utilizzando un *software* differente. Ciò ha comportato necessariamente l'avvio di una fase di test importante e attenta che ha riguardato:

- il *wording* di tutti i quesiti e dell'"*Help in linea*" (nonché l'effettivo rimando allo stesso nei quesiti interessati, ove indicato);
- l'eventuale presenza di errori di percorso e di flusso e quindi l'errata apertura dei quesiti;
- la corretta attivazione di regole e i rispettivi messaggi di *warning* visualizzati.

L'adozione della tecnica mista CATI e CAPI è stata affiancata dalla scelta di utilizzare un questionario unico, adatto sia per la somministrazione telefonica, sia per l'intervista diretta. Il nuovo questionario ha sostanzialmente recepito due grandi novità introdotte con la sperimentazione:

- le modifiche di *wording*, apportate per rendere l'intervista più fluida e aumentare la comprensibilità dei quesiti;
- lo snellimento del questionario, in virtù del maggiore utilizzo della base dati amministrativa in fase di trattamento dati.

Per quanto riguarda il primo punto, si è recepito lo sforzo fatto in sede di sperimentazione per ottenere quesiti più chiari e brevi, adatti anche alla lettura telefonica. Rispetto al secondo punto, è stata confermata la decisione di non rilevare per gli individui del campione teorico, le componenti di reddito desumibili da archivio amministrativo, come ad esempio le informazioni sulle pensioni.

Con la progettazione del nuovo questionario, inoltre, è stata affinata la strategia di rilevazione delle variabili di reddito. Infatti, la sperimentazione del "doppio rilancio", realizzata durante l'esperienza CATI e volta a cogliere, per alcune variabili reddituali, almeno un importo approssimativo, ha confermato l'utilità del secondo quesito in classi per recuperare una importante quota di informazione per tutti i quesiti per cui era già stato previsto (ad esempio per il reddito netto familiare) mentre si è rilevata inefficace in altri casi, rendendone superflua la somministrazione. Per tutte le altre variabili quantitative è stata confermata l'utilità del semplice sollecito del rilevatore, suggerito dal relativo *warning*.

Per quanto riguarda il contatto con le famiglie, nella sperimentazione era stata utilizzata la stessa sequenza di quesiti (detta Scheda Contatti) usata nelle precedenti edizioni dell'indagine, svoltesi in modalità CAPI. Tuttavia, data la peculiarità dei contatti telefonici, con la progettazione della nuova indagine mista CAPI/CATI la scheda è stata ridisegnata, con l'ottica di renderla adatta anche ad un contatto telefonico, e il nuovo formato è stato poi utilizzato per le interviste condotte con entrambe le tecniche. In particolare, nella nuova scheda, il contatto viene stabilito con i singoli individui della famiglia da intervistare e non in modo generico con la famiglia dell'intestatario ("Avrei bisogno di parlare con *tizio*" Anziché "È la famiglia di *tizio*?")

I rilevatori incaricati di svolgere la nuova indagine con tecnica mista sono stati formati secondo le modalità testate in occasione della sperimentazione CATI, ossia un corso dalla struttura snella e operativa con teoria e esercitazioni integrati in un'unica sessione. Il corso di formazione è stato unico per i rilevatori CATI e CAPI, ad eccezione del modulo relativo alla Scheda Contatti, dato che il contatto con la famiglia richiede strategie e accorgimenti diversi a seconda della tecnica. Nel realizzare la formazione si è tenuto conto degli aspetti critici emersi durante le attività di monitoraggio di sala e di *de-briefing* svolte nella sperimentazione CATI.

possibile effettuare un cambiamento di tecnica in corso di rilevazione, affidando la famiglia ad un rilevatore CAPI.

La nuova rilevazione si è svolta da giugno a dicembre 2016 per entrambe le tecniche; la rete CATI ha lavorato fisicamente in Puglia, dove era ubicato il *call center* della nuova Società incaricata. Come per la sperimentazione CATI, il monitoraggio della qualità da parte del personale Istat competente dell'indagine è stato realizzato sia attraverso l'analisi di indicatori specifici prodotti dalla Società incaricata, sia attraverso la supervisione in sala delle interviste.

Rispetto al sistema di indicatori di monitoraggio, per l'edizione 2016 è stato adottato un sistema semplificato rispetto a quello articolato e complesso progettato e utilizzato per la sperimentazione CATI.

Per quanto riguarda il monitoraggio di sala, visto che la sede del *call center* era lontano da Roma, è stato necessario ripensare e organizzare diversamente il monitoraggio della rete di rilevazione CATI. Le possibili soluzioni individuate sono state: il monitoraggio di sala nel *call center* in collaborazione con l'ufficio territoriale Istat della Puglia e il monitoraggio da remoto. Il coinvolgimento dei colleghi dell'ufficio territoriale ha reso possibile, già dal 2016, non solo un monitoraggio efficace, ma ha anche permesso di fornire ai rilevatori, mantenendo un costante contatto con il team EU-SILC a Roma, un valido supporto a eventuali problematiche riscontrate durante il periodo della rilevazione, tramite attività di supporto diretto in sala e di incontri di *de-briefing*. Inoltre, negli anni successivi, è stato reso possibile un monitoraggio dei rilevatori CATI da remoto, tramite l'ascolto non partecipante, nella sede romana della Società esterna incaricata. Le criticità emerse sono state condivise anche con la rete dei rilevatori CAPI, per i quali, l'attività di controllo della qualità, oltre che sugli indicatori di monitoraggio, si è basata anche su telefonate di controllo a campione.

2.4.2 La tecnica mista CAPI/CATI: i risultati della rilevazione dal 2016 al 2018

L'allocazione delle famiglie da intervistare tra le due tecniche, CATI e CAPI, ha dovuto necessariamente tenere conto sia della disponibilità di recapiti telefonici delle famiglie campione, sia dei tassi di risposta attesi, differenziati per tecnica e *wave*, in modo da conseguire il target prefissato in termini di numerosità effettiva del campione rilevato (circa 21.000 famiglie).

Per ciò che riguarda la disponibilità dei numeri di telefono, questa è stata piuttosto limitata (25-30% circa) per le famiglie campione alla loro prima intervista, estratte casualmente dalle Liste Anagrafiche Comunali (LAC) e per le quali i recapiti telefonici possono essere ricavati soltanto dagli elenchi pubblici. Tale disponibilità risulta invece superiore ma non completa (80-85% circa) per le famiglie già intervistate negli anni precedenti, quando si è provveduto, laddove possibile, a raccogliere direttamente al domicilio i recapiti telefonici da utilizzare nelle successive occasioni di indagine.

La scarsa disponibilità di numeri di telefono per il campione di prima *wave*, unitamente alle difficoltà di reperire numeri telefonici corretti durante le interviste e ai tassi di risposta specifici della tecnica, ha avuto come conseguenza che la percentuale di interviste CATI è diminuita negli anni, passando dal 57% nel 2016 al 43,7% nel 2018 (Tavola 2.4).

Tavola 2.4 - Famiglie intervistate per wave, anno di indagine e tecnica (valori assoluti, incidenza percentuale per tecnica)

WAVE	EU-SILC 2016			EU-SILC 2017			EU-SILC 2018		
	CAPI	CATI	%CATI	CAPI	CATI	%CATI	CAPI	CATI	%CATI
1	6.703	2.514	27,3	4.951	2.339	32,1	4.397	1.551	26,1
2+	2.163	10.037	82,3	5.303	9.692	64,6	7.525	7.701	50,6
Totale	**9.217**	**12.200**	**57,0**	**10.254**	**12.031**	**54,0**	**11.922**	**9.252**	**43,7**

L'indagine condotta con tecnica mista mostra prestazioni simili per il triennio 2016-2018. Per semplicità, facciamo riferimento ai dati più recenti disponibili (edizione 2018).

I tassi di completezza relativi alla componente CATI del 2018 sono migliori di quelli osservati durante la sperimentazione CATI (68,7% e 53,2% rispettivamente).

Ciò può essere stato determinato da due fattori concomitanti: in primo luogo, come accennato in precedenza, la partecipazione all'indagine pilota non era obbligatoria e questo può avere in quell'occasione sfavorito la partecipazione delle famiglie, soprattutto di quelle che erano state precedentemente intervistate per i quattro anni di indagine EU-SILC (tasso di completezza al 56,4% contro il 75% osservato nella componente CATI del 2018 per le *wave* diverse dalla prima); in secondo luogo, possono aver dato i loro frutti la messa a punto di strumenti di rilevazione più adeguati e la formazione mirata dei rilevatori.

Il tasso di completezza per tecnica mostra che la CAPI ha prestazioni migliori sia in prima *wave* (il 63,5% delle interviste complete rispetto al 48,4% in CATI), sia nelle *wave* successive (86,7% delle interviste complete in CAPI e 75% in CATI) (Tavola 2.5).

Per quanto riguarda le diverse componenti che caratterizzano la mancata risposta nelle due tecniche occorre distinguere tra le famiglie che sono alla loro prima intervista e quelle che ne hanno già affrontata almeno una negli anni precedenti.

Nella prima *wave*, il campione CAPI e quello CATI mostrano livelli simili dell'incidenza dei rifiuti (rispettivamente il 14,1% e il 16,9%). Le percentuali di non risposta CATI sono dovute soprattutto a numeri di telefono errati (16,5%) e ai tentativi di contatto senza risposta durante il lavoro sul campo (18,1%). Al contrario, la mancata risposta CAPI è dovuta in modo consistente alle difficoltà di copertura territoriale di questa tecnica nei tempi consentiti per lo svolgimento della rilevazione, con una percentuale di famiglie che al termine del *field* non fanno registrare alcun tentativo di contatto pari al 12,9%.

Nelle *wave* successive alla prima, le percentuali di risposta salgono sensibilmente per entrambe le tecniche, mentre la mancata risposta CATI dovuta a numeri errati resta ancora elevata (12,7%). Vale la pena notare che i numeri di telefono della prima *wave* sono forniti dai registri pubblici, mentre nelle *wave* successive si aggiungono i numeri di telefono raccolti, non sempre con adeguata precisione, durante le precedenti interviste.

Tavola 2.5 - Famiglie per esito dell'intervista, tecnica e wave, EU-SILC 2018 (composizione percentuale per esito e valori assoluti)

	CAPI			CATI			TOTALE
	Wave 1	Wave 2+	Totale	Wave 1	Wave 2+	Totale	
Interviste complete	63,5	86,7	76,4	48,4	75,0	68,7	72,8
Interviste incomplete:							
- rifiuto	*14,1*	*4,7*	*8,8*	*16,9*	*3,2*	*6,5*	*7,7*
- nome/indirizzo/telefono sconosciuto	*1,6*	*0,9*	*1,2*	*16,5*	*12,7*	*13,6*	*6,9*
- famiglie con soli tentativi di contatto a fine periodo di rilevazione	*4,2*	*0,8*	*2,3*	*18,1*	*8,7*	*11,0*	*6,3*
- famiglie senza tentativi di contatto a fine periodo di rilevazione	*12,9*	*5,0*	*8,5*	*0,1*	*0,1*	*0,1*	*4,6*
- altri motivi di caduta	*3,7*	*1,9*	*2,7*	*0,1*	*0,3*	*0,2*	*1,6*
Totale famiglie	**6.923**	**8.681**	**15.604**	**3.203**	**10.268**	**13.471**	**29.075**

La durata media di intervista è minore in CATI rispetto al CAPI (rispettivamente 25 minuti e 39 minuti) (Tavola 2.6), probabilmente perché al telefono può risultare meno agevole uscire dagli argomenti del questionario, cosa che invece, nelle interviste al domicilio, può essere desiderabile per stabilire un clima di collaborazione tra i rilevatori e i rispondenti e dove i rispondenti possono trovare agevole la consultazione di documenti utili a fornire valori esatti delle componenti del reddito o delle spese della famiglia.

Si noti che le durate medie sono più brevi nelle *wave* successive alla prima. Ciò può essere dovuto ad una maggiore dimestichezza dei rispondenti con i contenuti del questionario, acquisita con le interviste precedenti, ma anche all'efficacia dei quesiti a conferma: per alcune informazioni, come ad esempio le caratteristiche demografiche, il livello di istruzione e l'occupazione principale, agli intervistati viene chiesto solo di confermare le informazioni già raccolte negli anni precedenti (il resto del questionario viene somministrato normalmente). Ciò consente di ridurre il fastidio statistico sui rispondenti oltre che di controllare ed eventualmente correggere le informazioni precedentemente raccolte.

Tavola 2.6 - Durata media delle interviste (in minuti) per numero di componenti della famiglia, tecnica e wave EU-SILC 2018 (a)

NUMERO COMPONENTI	CAPI			CATI		
	Wave 1	Wave 2+	TOTALE	Wave 1	Wave 2+	TOTALE
1	35,1	29,0	31,1	21,0	17,5	18,0
2	47,7	40,1	43,1	30,4	23,4	24,5
3	55,7	47,1	50,4	35,5	30,2	31,2
4+	68,3	53,5	59,6	44,0	35,2	37,0
Totale	**44,1**	**35,9**	**38,9**	**30,3**	**23,9**	**25,0**

(a) La durata media dell'intervista è calcolata tenendo conto del tempo per completare la scheda famigliare, il questionario famigliare e tutti i questionari individuali.

2.5 Conclusioni

In linea con le raccomandazioni europee e con gli obiettivi nazionali, nell'indagine EU-SILC, tra il 2013 e il 2014, è stata sperimentata la tecnica CATI congiuntamente ad un maggiore sfruttamento delle informazioni provenienti dagli archivi amministrativi. Successivamente, a partire dall'edizione del 2016, sulla base dei risultati ottenuti e di vincoli organizzativi e di costo, è stata implementata a regime la tecnica mista CAPI/CATI.

L'implementazione della tecnica CATI nella rilevazione EU-SILC, oltre a consentire una netta riduzione dei costi complessivi di indagine, ha permesso di conseguire una serie di vantaggi già noti in letteratura: riduzione dei tempi di rilevazione; copertura adeguata anche delle zone meno accessibili del territorio; riduzione dell'invasività negli spazi domestici delle famiglie e riduzione del *response burden*; possibilità di supportare e monitorare l'attività dei rilevatori, riuniti in un'unica sede.

Si può ritenere che questo ultimo aspetto abbia contribuito in modo significativo al conseguimento di un'elevata qualità delle interviste: da un lato ha consentito di fornire ai rilevatori chiarimenti e di correggerli *in itinere* (anche tramite i *de-briefing*), dall'altra di rendere tangibile una supervisione sul loro lavoro. L'attività di monitoraggio, soprattutto nella fase di sperimentazione, ha inoltre fornito un ritorno prezioso, non solo sull'andamento della rilevazione dei dati, ma anche su altre fasi dell'indagine: progettazione del questionario, formazione, controllo, correzione, elaborazione dei dati, in quanto ha messo in luce alcune criticità sulla fluidità dei percorsi, su tematiche (e relative batterie di domande) che creano resistenza/imbarazzo/difficoltà negli intervistati e/o difficoltà per i rilevatori, su formulazione e *wording* di specifici quesiti. L'ascolto non partecipante ha anche permesso di evidenziare alcune criticità nella conduzione delle interviste da parte dei rilevatori CATI, che sono divenuti importanti punti da condividere anche alla rete di rilevatori CAPI e da sottolineare in sede formativa.

Sono risultate molto chiare anche alcune problematiche legate alla tecnica telefonica: i tassi di risposta sono più bassi che nelle interviste a domicilio; il ricorso alla risposta *proxy* è più frequente; l'accuratezza delle informazioni rilevate può essere inferiore anche laddove la presenza di tipologie familiari con più componenti o con situazioni reddituali articolate comporti una durata eccessiva dell'intervista e una maggiore difficoltà a mantenere alta l'attenzione dei rispondenti.

In generale, la possibilità di utilizzare la tecnica di rilevazione mista permette, tramite la combinazione delle due tecniche, di sfruttare al massimo le potenzialità e i vantaggi di entrambe e di minimizzarne gli svantaggi.

Del resto, la limitata disponibilità dei recapiti telefonici delle famiglie appartenenti al campione è un vincolo che, aggiunto a quello economico, limita di fatto la possibilità di distribuire il campione per tecnica secondo criteri ragionati, quali la *wave* di appartenenza e la potenziale complessità dell'intervista dovuta ad una elevata numerosità attesa di percettori e/o tipologie percepite di reddito.

Tra i vari aspetti da approfondire ulteriormente in futuro riguardo l'uso della tecnica mista, rimane la valutazione dell'impatto della tecnica sulle modalità di risposta fornite dagli intervistati. In altri termini occorrerà, con appropriate metodologie, indagare se uno stesso intervistato abbia propensione a rispondere in modo differenziato ad un medesimo quesito laddove esso sia posto in un'intervista telefonica piuttosto che in un'intervista diretta.

PARTE SECONDA

3. I METODI E LE TECNICHE PER IL CONTROLLO, LA CORREZIONE, L'IMPUTAZIONE E LA VALIDAZIONE DEI DATI[1]

3.1 Il controllo e la correzione dei dati

Come detto nei capitoli precedenti, a partire dal 2011, la rilevazione di EU-SILC è stata svolta tramite l'utilizzo di un questionario elettronico, che ha permesso di risolvere già al momento dell'intervista la gran parte degli errori che caratterizzavano l'acquisizione delle informazioni effettuata tramite questionario cartaceo e di ridurre i tempi e i costi del processo.

Fino al 2010, infatti, per ogni famiglia intervistata il rilevatore doveva compilare diversi modelli: una scheda generale con le informazioni anagrafiche di tutti i componenti, un questionario familiare e tanti questionari individuali quanti risultavano essere i componenti di almeno 15 anni della famiglia.

Lo svolgimento dell'intervista veniva gestito dal rilevatore osservando le prescrizioni stampate sul questionario: filtri sui quesiti o sulle sezioni del questionario, rimandi ad altri quesiti o sezioni condizionati dalle risposte ai quesiti precedenti.

Ciò comportava la possibilità di errori di "percorso", ovvero di rivolgere al rispondente dei quesiti non pertinenti o, al contrario, di omettere dei quesiti necessari, oltre alla possibilità di errori di compilazione o di classificazione delle risposte ottenute.

Una volta raccolti i modelli di rilevazione da tutti i rilevatori coinvolti, il comune li trasmetteva all'Istat dove si doveva provvedere ad una preliminare fase di spoglio manuale dei modelli per abbinarli correttamente per famiglia, accettarli o rigettarli in base al grado di completezza riscontrato, verificare ed eventualmente correggere almeno le principali variabili anagrafiche, conteggiarli e avviarli alla registrazione su supporto informatico, dove potevano sorgere ulteriori errori dovuti a inaccuratezza nella trascrizione[2].

A partire dal 2011, l'acquisizione delle informazioni in modalità assistita dal computer mette al riparo dalla gran parte degli errori sopra descritti perché il questionario elettronico gestisce in modo automatico il percorso dell'intervista e prevede una serie di controlli sui dati inseriti che possono segnalare eventuali inconsistenze sulle quali è possibile intervenire immediatamente, in presenza e in accordo con il rispondente.

Anche in questa modalità, tuttavia, la fase di controllo, correzione e imputazione dei dati rimane un'esigenza imprescindibile per l'indagine EU-SILC per diversi motivi:

1. non in tutti i casi è ragionevole inserire nel questionario elettronico dei controlli di consistenza vincolanti tra le informazioni acquisite nel corso dell'intervista, specialmente se queste sono raccolte in punti distanti del questionario stesso, e si è ritenuto preferibile demandare a una successiva fase di riconciliazione;

1 Il capitolo è stato redatto da Silvano Vitaletti.

2 La registrazione dei dati (affidata in service a una società esterna all'Istat) veniva effettuata in modalità controllata (CADI - Computer Assisted Data Inputing) per i codici identificativi e per le principali variabili anagrafiche. Erano inoltre previsti vincoli di dominio per le variabili qualitative (ammissibilità valori) e quantitative (controlli di *range*). Sul materiale proveniente dalla registrazione venivano condotti gli opportuni controlli statistici di qualità mediante confronto con un campione di documenti originali e il conteggio delle quantità registrate (Ceccarelli *et al.*, 2008).

2. i dati rilevati tramite intervista vengono successivamente integrati con una molteplicità di dati di fonte amministrativa, determinando possibili inconsistenze che devono essere riconciliate;
3. per le variabili di reddito viene ammessa la possibilità di non rispondere, ritenendo preferibile un'omissione, recuperabile tramite imputazione e integrazione con i dati amministrativi, alla rilevazione di informazioni falsate dalla reticenza dei rispondenti;
4. anche quando veritiere e plausibili, le informazioni possono essere statisticamente non conformi alla distribuzione complessiva (dati anomali/*outliers*) e devono essere opportunamente trattate per evitare di influenzare in modo sostanziale le stime;
5. i dati acquisiti in occasioni di indagine diverse per la medesima unità di rilevazione possono dar luogo a incoerenze che devono essere opportunamente gestite in fase di trattamento dati.

3.2 Le fasi del processo di controllo e correzione dei dati

Il processo di controllo e correzione dei dati si articola sequenzialmente nelle fasi di seguito sommariamente descritte.

Correzione delle variabili strutturali individuali

Vengono congiuntamente controllate e corrette in modo deterministico le chiavi individuali, affette da eventuali duplicazioni o inversioni, e alcune variabili identificative (nome, cognome, sesso, data e luogo di nascita, ecc.), affette da eventuali errori di rilevazione o registrazione.

Dal 2011, con l'introduzione della richiesta di conferma sulle variabili identificative, precaricate nel questionario elettronico, si è drasticamente ridotto l'impatto di questa procedura.

Data la natura longitudinale dell'indagine, è fondamentale nella fase di controllo e correzione delle chiavi individuali e familiari assicurare il rispetto della coerenza longitudinale. La procedura generalizzata adottata viene descritta nell'ambito della quarta parte del presente volume, nel Capitolo 14 "La trasformazione dei file provenienti dalla rilevazione alla base dei processi di trattamento dei dati dell'indagine".

Correzione delle variabili strutturali familiari

Le eventuali inconsistenze tra alcune variabili demografiche che definiscono la struttura familiare (sesso, età, relazione di parentela con la persona di riferimento della famiglia, stato civile) vengono individuate e corrette, mediante il software generalizzato noto come "procedura famiglie" (Budano e Demofonti, 2010), il quale identifica anche le coppie e le tipologie familiari.

Il passaggio da PAPI a CAPI ha determinato una drastica riduzione dell'impatto di questa procedura: considerando il numero totale di incroci unità-variabile, la percentuale di casi toccati è passato dal 20,1% del 2010 al 6,0% del 2013.

Integrazione dei dati rilevati con l'intervista con i dati provenienti da fonti amministrative – Identificazione dei percettori di reddito

I dati provenienti dalle interviste vengono abbinati alle informazioni contenute negli archivi amministrativi (modelli di dichiarazione fiscale CUD/730/UPF dall'Agenzia delle Entrate, casellario pensionistico e altri archivi su trasferimenti sociali non pensionistici dall'INPS) con tecniche di *record linkage* esatto mediante codice fiscale, ricavato direttamente dalle Liste Anagrafiche Comunali o ricostruito mediante procedura ad hoc, in en-

trambi i casi validato dalla SOGEI anche sulla base di alcune informazioni ausiliarie (indirizzo della residenza)[3].

Il primo passo di integrazione consiste nell'identificazione dei percettori delle varie tipologie di reddito riconciliando l'informazione rilevata durante l'intervista con quella risultante dai registri amministrativi.

Al termine di questo passo vengono fissate delle variabili indicatrici che informano, per ciascun individuo, se esso ha percepito una o più delle diverse tipologie di reddito previste.

Correzione e imputazione delle variabili qualitative

Per la correzione e l'imputazione automatica delle variabili qualitative viene impiegata la metodologia Fellegi-Holt (Fellegi e Holt, 1976) implementata nel software generalizzato Concord (SCIA) sviluppato in Istat (Istat, 2004). Tale metodologia è basata sui seguenti principi:

- Minimo cambiamento dei dati. Viene identificato l'insieme minimo di variabili che necessita di essere corretto all'interno dello stesso record per risolvere un'eventuale incompatibilità.
- Correzione e imputazione come processo unico. L'insieme completo delle regole di incompatibilità (*edit*), oltre a consentire la separazione dei record errati da quelli esatti, definisce una regione di ammissibilità dalla quale vengono prelevati i valori imputati che, in questo modo, soddisfano con certezza tutte le regole.
- Mantenimento della struttura dei dati. Vengono il più possibile conservate le distribuzioni marginali e congiunte del sottoinsieme dei dati originari nel quale non sono presenti incompatibilità.

L'insieme completo di regole di incompatibilità (relazioni non ammissibili tra valori di variabili diverse), è composto da "regole formali" e "regole sostanziali".

Le prime sono definite dalla struttura stessa del questionario e dalle norme di compilazione: una determinata risposta a un quesito rimanda a un altro quesito, non necessariamente il successivo.

Le regole sostanziali identificano, viceversa, altre incompatibilità non definite dai possibili percorsi del questionario e che derivano da conoscenze specifiche a priori del fenomeno rilevato. È il caso, per esempio, dell'impossibilità di avere, per una famiglia senza bambini, uno dei componenti che riceve benefici legati alla presenza di bambini, quale un assegno di maternità o un assegno al nucleo con almeno tre figli. Particolarmente rilevanti sono le regole che collegano le variabili derivanti dall'integrazione con i dati amministrativi che identificano il tipo di reddito percepito e le altre variabili dell'indagine. Nel caso, ad esempio, di un percettore di redditi da pensione da lavoro (da casellario pensionistico) tale informazione è incompatibile con la dichiarazione (in sede di intervista) di non avere mai lavorato in passato. Alcune regole, inoltre, basandosi sull'impianto longitudinale dell'indagine, legano le informazioni rilevate dall'intervista attuale con quelle rilevate l'anno precedente: il livello di istruzione attuale, ad esempio, non può essere inferiore a quello rilevato nella precedente intervista.

In ogni caso, nel processo di controllo e correzione devono necessariamente essere incluse anche le regole formali, in quanto in caso di imputazione di nuovi valori per rimuovere incoerenze sostanziali o valori mancanti, essa deve avvenire inevitabilmente nel rispetto delle norme di compilazione del questionario.

Nell'ambito del processo di EU-SILC, le regole formali vengono generate automaticamente tramite il Sistema Automatico di Generazione Edit (SAGE), che agisce utilizzando la formalizzazione del questionario come grafo aciclico diretto, il quale viene esplorato per

3 Consolini, 2009.

individuare gli "archi/percorsi" e i "nodi/variabili" incidenti al fine di derivare le regole di incompatibilità. Tale metodo e la sua combinazione con un approccio ragionato per la definizione dell'intero piano di compatibilità tra le variabili di indagine viene descritto in modo più approfondito nel Capitolo 4 "SAGE: Il Sistema automatico di generazione delle regole di incompatibilità nelle variabili del questionario".

Occorre sottolineare che l'adozione del questionario elettronico a partire dal 2011 ha annullato quasi del tutto l'insorgenza di incompatibilità "formali" nei dati grezzi, eccettuati i casi di risposte mancanti, ammessi solo per le interviste *proxy* e per alcuni particolari quesiti, soprattutto relativi ad informazioni di tipo quantitativo. Anche l'insorgenza delle incompatibilità "sostanziali" è stata ridotta con il questionario elettronico tramite l'implementazione di numerosi controlli (circa 160 di tipo "soft", 20 di tipo "hard"[4]) direttamente in corso di intervista.

Non sono sicuramente gestibili, in fase di intervista, le incompatibilità generate dall'integrazione con i dati amministrativi, che avviene in un momento successivo. Pertanto, nell'ambito della correzione probabilistica, tali variabili vengono mantenute fisse (nel limite delle possibilità), preferendo la correzione/imputazione delle rimanenti variabili del questionario, qualora sorgano delle incompatibilità con la tipologia di reddito di fonte amministrativa.

Più in generale, data la complessità del questionario di EU-SILC e la numerosità delle regole e delle variabili implicate, si è reso necessario sezionare gerarchicamente il processo in più passi:

1. Scheda generale;
2. Questionario familiare;
3. Stato di occupazione attuale e in passato;
4. Prospetto redditi (percepimento delle diverse tipologie di reddito);
5. Situazione relativa al lavoro o alla ricerca di lavoro;
6. Reddito da lavoro dipendente;
7. Reddito da lavoro autonomo;
8. Reddito da pensioni;
9. Reddito da altre fonti;
10. Istruzione e condizioni di salute.

Tale organizzazione gerarchica dei passi di correzione impone che le variabili corrette/imputate in un passo vengano considerate fisse, quindi non più modificabili, se vengono coinvolte in regole nei passi successivi.

Come anticipato, l'introduzione del questionario elettronico ha determinato una riduzione dell'impatto complessivo della procedura di correzione e imputazione probabilistica delle variabili qualitative. Tenendo conto di tutte le celle di incrocio tra unità e variabili, la percentuale di quelle trattate sul totale, infatti, si è quasi dimezzata, passando dal 6,1% del 2010 al 3,5% del 2013. Gli interventi sulle variabili della Scheda e del Questionario familiare attualmente sono minimi, riguardano infatti rispettivamente solo lo 0,5% e lo 0,2% delle unità-variabili, mentre le correzioni e imputazioni interessano in maggior misura le variabili del Questionario individuale (5,1% delle unità-variabili) trattandosi del modello di rilevazione dove sono rilevate la maggior parte delle tipologie di reddito e le situazioni lavorative individuali.

4 I controlli di tipo "soft" sono quelli che forniscono un'avvertenza per il rilevatore al fine di verificare la correttezza dei valori immessi, anche mediante confronto con l'intervistato, dove la verifica può portare anche ad una conferma di quanto inserito inizialmente. I controlli di tipo "hard" non sono invece aggirabili da parte dell'intervistatore in quanto riguardano l'individuazione di una situazione non ammissibile.

Tavola 3.1 - Celle (a) **trattate nel processo di correzione/imputazione delle variabili qualitative per tipo di trattamento e modello di rilevazione** (valori percentuali)

Tipo di trattamento	Scheda generale	Questionario familiare	Questionario individuale	Totale
		IT-SILC 2013 (CAPI)		
Immutate	94.7	99.0	92.1	93.9
Modificate	0.9	0.1	1.0	0.8
Imputate	3.7	0.4	5.8	4.3
Sbiancate	0.7	0.5	1.1	0.9
Totale	100.0	100.0	100.0	100.0
		IT-SILC 2010 (PAPI)		
Immutate	99.5	99.8	94.9	96.5
Modificate	0.2	0.0	0.8	0.6
Imputate	0.1	0.0	3.6	2.5
Sbiancate	0.2	0.2	0.6	0.5
Totale	100.0	100.0	100.0	100.0

(a) Ogni cella corrisponde a una singola unità campionaria e a una singola variabile. Sono considerate soltanto le variabili qualitative.

Imputazione da donatore dei questionari individuali mancanti

Questa fase di correzione si rende necessaria qualora, all'interno di una famiglia rispondente, ci siano uno o più individui di almeno 16 anni (eleggibili per l'intervista personale) non rispondenti: questa casistica, individuata come '*partial unit non-response*' dalle linee guida Eurostat (European Commission, 2017, pag.54), in base alle indicazioni europee, può essere gestita mediante un intervento sui coefficienti di riporto all'universo individuali oppure attraverso un'operazione di imputazione. In Italia si è optato per l'imputazione, anche in considerazione dell'esiguo insieme dei casi da trattare.

Il metodo adoperato è quello del donatore di distanza minima, implementato nella procedura *Full Record Imputation* (FRI), sviluppata internamente al processo di trattamento dati di EU-SILC. La procedura in questione viene descritta in modo più approfondito nel Capitolo 5 "FRI: Una nuova procedura per l'imputazione da donatore".

In sintesi, prima è definito l'insieme dei possibili donatori, cioè gli individui rispondenti che appartengono allo stesso strato dell'individuo non rispondente. Lo strato è definito dall'intersezione della dimensione familiare, area geografica, classe di età, sesso, cittadinanza e stato di nascita (Italia/estero), stato civile, capacità della famiglia di arrivare a fine mese, titolo di godimento dell'abitazione.

Il donatore è selezionato, tra i potenziali record dello stesso strato del ricevente, in base alla minima distanza dal ricevente nello spazio identificato dalle variabili di *matching* (il reddito totale familiare).

Dal record del donatore vengono prese tutte le informazioni rilevate, ad eccezione di quelle sui redditi che vengono imputate successivamente. Il processo può essere reiterato, rilassando di volta in volta le condizioni che definiscono l'insieme dei possibili donatori, fino all'imputazione di tutti i record mancanti.

Correzione e imputazione delle variabili quantitative (redditi e spese per l'abitazione)

Prima di tutto si procede all'identificazione e rimozione dei valori anomali mediante il metodo Hidiroglou-Berthelot che consiste nella costruzione di un intervallo di accettazione basato sulle distanze inter-quartiliche della distribuzione univariata (Ceccarelli, Di Marco e Rinaldelli, 2008).

Per alcune variabili viene adottato un approccio bivariato basato sui rapporti di due variabili. Le spese per l'abitazione, ad esempio, sono considerate in rapporto alle dimensioni dell'alloggio (superficie e numero di stanze), i redditi finanziari sono valutati in rapporto al volume complessivo dei risparmi.

Per le informazioni rilevate nelle wave successive alla prima, il metodo è applicato in ottica longitudinale, considerando il rapporto tra il valore corrente e quello rilevato nella precedente occasione di indagine.

Per l'imputazione dei valori mancanti, ivi compresi quelli generati dalla rimozione dei valori anomali, è impiegato il metodo di regressione multivariata sequenziale (implementato nel software IVEware dell'University of Michigan)[5]:

- l'imputazione è effettuata in modo sequenziale, variabile per variabile, a partire da quella con il minor numero di valori mancanti;
- sono utilizzati modelli di regressione differenti a seconda del tipo di variabile (continuo, dicotomico, conteggio, ecc.);
- l'imputazione è attuata secondo uno schema iterativo, in cui viene aggiornato l'insieme delle covariate in funzione delle imputazioni precedenti, tendendo a preservare la correlazione tra le variabili di volta in volta soggette a imputazione.

I modelli sfruttano come covariate sia le informazioni disponibili a priori, come alcune caratteristiche territoriali (area geografica, tipologia di comune), sia informazioni già validate nel processo di correzione e imputazione delle variabili qualitative. Si tratta di informazioni a livello sia familiare (dimensione della famiglia, caratteristiche dell'abitazione, possesso di beni durevoli, indicatori di disagio economico, ecc.), sia individuale (sesso, età, stato civile, cittadinanza, istruzione, condizioni di salute, caratteristiche dell'occupazione, ecc.).

Per alcune variabili sono state definite delle restrizioni, consentendo l'imputazione solo su sottoinsiemi definiti di casi, e/o dei vincoli di dominio, consentendo ai valori imputati di variare tra un minimo e/o un massimo opportunamente definiti.

Per poche variabili, come ad esempio i redditi da lavoro autonomo o i redditi da capitale immobiliare, in caso di mancata risposta al valore puntuale, è prevista la possibilità di rilevare la classe di importo approssimato cui il valore effettivo appartiene. In tali casi il dominio di imputazione viene definito dai limiti inferiore e superiore della classe. In generale, per tutte le variabili che non prevedono la rilevazione degli importi in classi o quando tale opzione non sia stata scelta dal rispondente, il dominio di imputazione viene definito dall'intervallo di accettazione calcolato dalla procedura di individuazione dei valori anomali di cui sopra. Per alcune tipologie di trasferimenti sociali il dominio di imputazione è definito dalla normativa vigente.

La procedura di imputazione affronta inizialmente le spese per la casa, i redditi e i trasferimenti a livello familiare, in quanto aiutano a stabilire il tenore di vita della famiglia e quindi a spiegare il livello delle altre componenti di reddito. Successivamente vengono affrontate le altre sezioni, a partire da quelle con la minore incidenza di valori mancanti. In definitiva la sequenza del processo è così articolata:

1. spese per la casa, redditi e trasferimenti a livello familiare;
2. redditi da lavoro dipendente;
3. redditi da pensione;
4. redditi da lavoro autonomo;
5. rendite da capitale finanziario;
6. rendite da capitale immobiliare;
7. altri redditi;
8. altre informazioni quantitative sulla storia lavorativa.

Nel processo di correzione e imputazione delle variabili di reddito interviene in modo massiccio l'utilizzo dei dati amministrativi (Consolini, 2009). In termini molto

5 Ceccarelli *et al.*, 2008.

sintetici sono due le modalità principali secondo le quali il dato amministrativo viene riconciliato con il dato rilevato tramite intervista diretta. Il primo riguarda i redditi da lavoro autonomo, il secondo tutte le altre tipologie di reddito.

Nel caso dei redditi da lavoro autonomo la riconciliazione è effettuata dopo la correzione e l'imputazione dei dati delle interviste. Per ciascun soggetto, il valore ottenuto viene confrontato con il corrispondente valore di fonte amministrativa e il valore finale corrisponde al massimo tra i due. L'ipotesi sottostante è che entrambi i valori siano affetti da sottostima e scegliendo il massimo tra i due si riduce l'entità della sottostima stessa.

Per tutte le altre tipologie di reddito, il dato amministrativo è dapprima utilizzato per colmare le mancate risposte parziali all'intervista (o quelle dovute alla rimozione dei valori anomali). I residui valori mancanti vengono imputati con il metodo precedentemente descritto.

Al termine della procedura di imputazione vengono applicati dei correttivi sulle code delle singole distribuzioni per ridurre l'impatto sulle stime, specialmente su domini territoriali subnazionali, di dati potenzialmente influenti: quando un valore estremo (minore del 5° percentile o superiore al 95° percentile della distribuzione regionale), si associa a un peso molto elevato (oltre l'80° percentile della distribuzione regionale dei pesi), esso viene ricollocato in modo casuale rimanendo nella medesima metà della distribuzione rispetto alla mediana (tra il 5° e il 10° percentile oppure tra il 90° e il 95° percentile).

Transcodifica delle variabili target relative ai redditi netti

Le variabili di base relative ai redditi netti, corrette e/o imputate nel processo sopra esposto, vengono trasformate nelle variabili *target* richieste dal regolamento tramite istruzioni rigidamente deterministiche.

L'impatto della procedura di imputazione delle variabili di reddito viene quantificato mediante gli *imputation factor*, che sono a loro volta variabili incluse nei microdati finali. Per ciascuna variabile *target* di reddito, l'*imputation factor* è dato dal rapporto tra il valore iniziale prima del trattamento e il valore finale. Poiché una singola variabile *target* può essere somma algebrica di numerose variabili di base, la determinazione del suo valore iniziale è di una certa complessità soprattutto in presenza di valori mancanti. In ogni caso, qualora il valore finale della variabile di base sia di fonte amministrativa, il valore prima del trattamento, considerato ai fini dell'*imputation factor*, è quello amministrativo, in quanto dato "raccolto", sebbene da fonte alternativa all'intervista. Analogamente, qualora il valore imputato sia dovuto alla mera attribuzione di un valore puntuale in caso di acquisizione di una classe approssimata di importo in sede di intervista, il dato prima del trattamento, considerato per il computo dell'*imputation factor*, è quello esito della trasformazione dalla classe in valore puntuale.

Stima dei redditi lordi tramite integrazione con dati fiscali e modello di micro-simulazione

Le variabili *target* relative ai redditi lordi vengono calcolate sommando al valore delle *target* relative ai redditi netti l'importo delle tasse e dei contributi sociali obbligatori versati da ciascun soggetto. Nella maggioranza dei casi, le tasse vengono ricavate dagli archivi dei modelli di dichiarazione fiscale. Fanno eccezione i casi di mancato *linkage* con tali archivi, nei quali l'importo delle tasse viene ricostruito, sotto l'ipotesi di piena conformità alla normativa vigente, tramite modello di micro-simulazione (SM2)[6].

In tutti i casi l'importo dei contributi sociali obbligatori versati viene ricavato dal modello di micro-simulazione.

6 Donatiello, 2011.

Nel Capitolo 6 "La stima delle tasse e dei contributi sociali" sono illustrate alcune innovazioni introdotte per il miglioramento della qualità delle stime.

Transcodifica delle rimanenti variabili target trasversali, controllo e correzione longitudinale e creazione dell'archivio riconciliato di indagine trasversale e longitudinale

Al termine di tutte le fasi di trattamento dei dati, vengono create le variabili *target* richieste dal regolamento anche per le informazioni di tipo qualitativo, sempre con criteri deterministici.

Fino al 2013 è stata mantenuta una formale divisione tra il database trasversale e quello longitudinale di indagine. La produzione dei file longitudinali era successiva[7] a quella dei file trasversali e le eventuali correzioni indotte dall'analisi di coerenza longitudinale non necessariamente si ripercuotevano sulla componente trasversale. In realtà, come accennato, quasi ciascuno dei passi di trattamento incorpora anche degli elementi di sfruttamento delle informazioni longitudinali e si può quindi ritenere che, almeno a livello macro, sia sempre stata rispettata una sostanziale coerenza tra i dati trasversali e quelli longitudinali. Dal 2014, i due tipi di archivi sono stati fusi in uno solo, riconciliato, nel quale viene garantita coerenza anche a livello micro. A tal fine, Eurostat ha sviluppato e reso disponibile un'articolata procedura per verificare la presenza di errori nei dati sia trasversali sia longitudinali. Ai criteri di Eurostat, applicabili esclusivamente alle variabili *target*, se ne aggiungono molti altri, sviluppati internamente, che invece sono applicabili alle variabili di base dell'indagine nazionale.

Gli aspetti specifici riguardanti la componente longitudinale dell'indagine vengono trattati con maggiore dettaglio nella quarta parte del volume.

Validazione dei dati

Come fase conclusiva del processo di trattamento dei dati di indagine una particolare attenzione viene dedicata alla validazione dei principali risultati ottenuti, anche attraverso il confronto con fonti esterne per poter certificare la qualità dei dati e per segnalare eventuali criticità nel processo di trattamento e di stima, come descritto nel Capitolo 2.5 "Il processo di validazione con fonti esterne a supporto del trattamento dei dati".

7 La stessa tempistica prevista dal regolamento concede 5 mesi in più per la consegna della componente longitudinale: marzo t+2 anziché novembre t+1. C'è da notare che il regolamento del 2003 non è stato ancora modificato ma in virtù di un "*gentlemen's agreement*", dal 2014 i paesi si sono impegnati a consegnare a giugno t+1 sia la componente trasversale sia quella longitudinale.

4. SAGE: IL SISTEMA AUTOMATICO DI GENERAZIONE DELLE REGOLE DI INCOMPATIBILITÀ NELLE VARIABILI DEL QUESTIONARIO[1]

4.1 Introduzione

La complessità del processo di controllo e correzione dei dati dell'indagine EU-SILC è tale da richiedere approcci il più possibile automatizzati, che permettano di ridurre il tempo necessario per completare una fase di trattamento e il numero di persone necessarie per portarla a termine. Sono automatizzabili quelle fasi del processo di correzione che possono essere scomposte in "passi", per i quali siano individuabili dei metodi e delle funzioni matematiche per la correzione dei dati, in modo da realizzare degli algoritmi e quindi un *software* che possa essere eseguito agevolmente anche da una sola persona.

L'automazione e la standardizzazione dei trattamenti, inoltre, permette di ridurre la possibilità di generare errori sui dati mantenendo elevata la probabilità di individuare gli errori. Per quanto riguarda il processo di produzione dati dell'indagine EU-SILC, una fase molto delicata riguarda la correzione probabilistica di alcune variabili qualitative del questionario. Si vedrà in seguito che tale fase prevede la scrittura di regole formali che coinvolgono le variabili rilevate tramite il questionario, legando i valori che una variabile può assumere ai valori assunti da una o più delle altre variabili.

A questo scopo, è stato ormai da diversi anni testato ed implementato in EU-SILC un approccio basato sull'applicazione della Teoria dei Grafi (Massoli, 2008). Tale approccio è molto generale ed il suo scopo è decisamente più ampio di quello qui presentato. Infatti, formalizzando il questionario come un grafo, è possibile sfruttare tutte le proprietà matematiche dei grafi. In particolare, si possono enumerare ed analizzare tutti i percorsi possibili del questionario, in modo da controllare se i dati rilevati sono coerenti con la struttura del questionario stesso. L'enumerazione fornisce anche una misura della complessità del questionario e, nei casi in cui il numero dei percorsi possibili non sia elevato, una visualizzazione grafica dei percorsi del grafo, fornendo uno strumento valido per il supporto alle decisioni durante la fase di progettazione.

È importante sottolineare che, poiché i percorsi del questionario rappresentano dei legami tra le sue variabili, le regole formali della fase di correzione sono direttamente ottenibili dalla struttura matematica del grafo. Inoltre, sfruttando gli algoritmi di partizionamento dei grafi, è possibile ricavare le sezioni di *taglio minimo* del questionario, ovvero quelle parti che sono collegate alle altre tramite un solo percorso. Questo aspetto assume una certa rilevanza in tutti quei casi in cui la complessità del questionario richiede di procedere alla correzione per passi.

Si è realizzato dunque un *software* che implementa tutte le funzionalità suddette in un'ottica di automazione, sfruttando i metadati del questionario: a) tipo e dominio delle variabili di interesse; b) collegamenti diretti tra coppie di variabili in funzione delle modalità da esse assunte. Tali informazioni sono state raccolte in una comoda e generale forma tabel-

1 I paragrafi 4.1, 4.2, 4.3, 4.8 sono stati redatti da Pierpaolo Massoli; i paragrafi 4.4, 4.5, 4.6, 4.7 sono stati redatti da Daniela Lo Castro.

lare che funge da sorgente esterna per tutte le applicazioni del *software*. Tale approccio si è rivelato robusto, rimanendo uno strumento efficace anche dopo il passaggio di EU-SILC dalla tecnica di rilevazione PAPI a quella CAPI e CATI.

Questo capitolo si concentra sulla funzionalità di generazione automatica delle regole formali del questionario per la costruzione gerarchica dei passi di correzione probabilistica delle variabili qualitative dell'indagine. Il paragrafo 4.2, riportando alcune semplici definizioni della teoria, introduce alla trattazione matematica dell'approccio proposto. Il paragrafo 4.3 presenta i metadati necessari per la costruzione della struttura matematica del questionario e la struttura dei *data set* che raccolgono i dati rilevati. Il paragrafo 4.4 descrive come vengono generati gli *edit* formali, a partire dalla struttura matematica del grafo. Il paragrafo 4.5 presenta alcuni cenni teorici sulla correzione probabilistica e come questa viene impiegata nell'indagine EU-SILC. Nel paragrafo 4.6 si riporta un esempio di generazione automatica degli *edit* formali nell'ambito di un'occasione di indagine, mentre nel paragrafo 4.7 si mette a confronto l'approccio automatico con uno basato sulle considerazioni degli esperti dell'indagine. Alcune considerazioni conclusive sono discusse nel paragrafo 4.8.

4.2 Alcune definizioni e aspetti teorici

Un grafo è una coppia di insiemi $G=(V,E)$ in cui $V=\{v_1,v_2,\ldots,v_N\}$ è l'insieme dei vertici (o nodi) e $E \subset V \times V$ è l'insieme degli archi (o lati) $E=\{e_1,e_2,\ldots,e_M\}$ dove ogni arco è una coppia di vertici $e_h=(v_i,v_j)$. Due vertici sono adiacenti se esiste un arco che li congiunge. Due vertici incidenti su un arco sono detti estremità. In un grafo diretto gli archi sono delle coppie ordinate di vertici ($j>i$ se v_i precede v_j). Un percorso da v_i a v_k è una sequenza di vertici e archi in cui tutti i vertici sono distinti. Il numero di archi in un percorso è detto lunghezza del percorso. Un grafo che contiene solo percorsi è detto aciclico. Un percorso che non è contenuto in nessun altro percorso si chiama massimale. È sempre possibile inserire un unico vertice, detto sorgente, dal quale possono solo uscire uno o più archi e un unico vertice, detto pozzo, in cui possono solo entrare uno o più archi, in modo che ogni massimale abbia la sorgente ed il pozzo come estremità.

Sotto opportune ipotesi, un questionario si può rappresentare come un grafo diretto aciclico (DAG) con una sorgente ed un pozzo. Per poter trattare operativamente il grafo con N vertici e M archi, assume una particolare rilevanza la matrice di incidenza $\mathbf{I}=\{i_{ij}\}$ di dimensioni ($N\times M$) così definita:

$$i_{ij} = \begin{cases} +1 & se\ v_i \in e_j\ con\ e_j\ uscente\ da\ v_i \\ 0 & se\ v_i\ non\ è\ incidente\ su\ e_j \\ -1 & se\ v_i \in e_j\ con\ e_j\ entrante\ in\ v_i \end{cases}$$

Vale la relazione $\mathbf{II}^T=\mathbf{A}+\mathbf{D}$ dove $A=\{a_{ij}\}$ è la matrice (simmetrica) di adiacenza ($N\times N$) con elementi $a_{ij}=1$ se i vertici v_i e v_j sono adiacenti, $a_{ij}=0$ altrimenti. La matrice diagonale $\mathbf{D}=\{d_{ii}\}$, delle stesse dimensioni di A, contiene i gradi di ciascun vertice[2]. Queste matrici caratterizzano completamente il grafo e quindi la struttura del questionario. Esse sono fondamentali per l'applicazione di tutti gli algoritmi che il sistema qui presentato adotta. È importante

2 Si definisce grado d_{ii} del vertice v_i il numero di archi incidenti su v_i.

sottolineare che la struttura particolare del grafo diretto aciclico, a cui si può ricondurre la struttura di un questionario, rende particolarmente semplice l'applicazione degli algoritmi che verranno successivamente descritti.

4.3 I metadati del questionario

Supponiamo, senza perdita di generalità, che ciascuna intervista sia rappresentata da un record in uno solo o in diversi *data set*, corrispondenti a diverse sezioni del questionario (come accade per l'indagine EU-SILC). I valori assunti dalle variabili presenti in tali *data set* corrispondono alle risposte registrate durante la somministrazione dell'intervista.

Ci sono risposte che possono essere memorizzate in una variabile singola e risposte a scelta multipla che necessitano di più variabili.

La tipologia delle variabili è sia numerica sia testuale. Queste tipologie, rispettivamente, sono usate per registrare risposte di tipo categorico (genere dell'individuo, "si/no", titolo di studio, ecc.), risposte di tipo quantitativo (importi monetari, spese, redditi, ecc.) e risposte testuali (note, professione dell'individuo, ecc.). Per ciascuna variabile viene indicato un dominio o un intervallo, a seconda che sia qualitativa o quantitativa.

Queste informazioni possono essere facilmente memorizzate in un file cui accedono tutte le funzioni del *software*. In esso sono inseriti tutti i "salti" che fanno passare da una variabile ad un'altra a seconda della risposta data alla domanda memorizzata nella prima variabile. Nel sistema che viene qui descritto tale strumento è chiamato *routing file*. La struttura del *routing file*, contenente i metadati per la costruzione del grafo DAG, è riportata in Tavola 4.1.

Tavola 4.1 - Struttura del *routing file*

COLONNA	DESCRIZIONE	TIPO
x_1	Nome della variabile	carattere
x_2	Numero delle modalità	intero [0,+∞)
x_3	Valore minimo della variabile	intero (-∞,+∞)
x_4	Valore massimo della variabile	intero (-∞,+∞)
x_5	Gruppo della variabile	intero (-∞,+∞)
x_6	Numero di salti dalla variabile	intero [-1, 98]
x_7	Modalità per cui si ha il salto	intero (-∞,+∞)
x_8	Nome della variabile di arrivo	carattere

Il generico record del file individua una variabile del questionario riportata con il suo nome (nella colonna 1), le sue caratteristiche (nelle colonne da 2 a 5), e i percorsi che da questa possono scaturire (nelle colonne da 6 a 8), a seconda delle modalità che può assumere. Ogni variabile del *routing file* così riportata è un vertice del grafo. Tutti i salti (archi) sono scritti nel file in modo che la variabile a cui rimanda il salto sia rappresentata, nel grafo aciclico, da un vertice *successore* del vertice corrispondente alla variabile da cui parte il salto. Le variabili del file sono ordinate seguendo l'ordine delle domande del questionario.

Sono utilizzate variabili di tipo numerico per tutte quelle domande le cui risposte possono essere codificate con numeri progressivi (discreti o continui), mentre sono utilizzate variabili testuali per codificare domande a risposta aperta che memorizzano una parola o una frase. Per distinguere una variabile testuale da una numerica, si pone il numero di modalità pari a 0, il valore minimo pari al valore nullo e il valore massimo pari a 1. Per le variabili numeriche quantitative il numero di modalità viene posto uguale a 1, mentre i corrispondenti valori minimo e massimo costituiscono l'intervallo dei valori ammissibili. Per

le variabili numeriche qualitative, il numero di modalità viene posto maggiore di 0, mentre i corrispondenti valori minimo e massimo delimitano il dominio della variabile.

Inoltre, due o più variabili possono essere raccolte in gruppi logici di tipo *OR*, quando fanno riferimento a domande a risposta multipla (per ciascuna possibile scelta viene definita una variabile distinta). Nel caso di domande in cui una possibile risposta è mutuamente esclusiva rispetto ad altre (es. l'intervistato fornisce l'importo e in tal caso si valorizza una variabile quantitativa oppure dichiara di non conoscerlo e in tal caso si valorizza una variabile qualitativa per memorizzare l'informazione "non sa"), le variabili corrispondenti sono messe in gruppi logici *XOR* (*OR* esclusivo). Nel primo caso, l'etichetta del gruppo è un numero intero positivo progressivo mentre nel secondo si utilizza un numero intero negativo progressivo. In tutti i casi di variabili non appartenenti a nessun gruppo logico l'etichetta del gruppo è 0.

In molti casi, il flusso dell'intervista si basa sull'utilizzo di informazioni precaricate oppure si articola secondo alcune informazioni rilevate in sezioni del questionario distanti dalla posizione corrente. Affinché la struttura matematica del questionario sia un grafo DAG, occorre gestire tali situazioni inserendo nel *routing file* delle variabili ausiliarie. Tali variabili sono calcolate a partire dai valori delle variabili di base. Le informazioni necessarie per la costruzione delle variabili ausiliarie sono contenute in un secondo *data set* detto *auxiliary file* (Tavola 4.2).

Tavola 4.2 - Struttura dell'*auxiliary file*

COLONNA	DESCRIZIONE	TIPO
Y_1	Nome della variabile ausiliaria	carattere
Y_2	Modalità della variabile ausiliaria	intero [1,+∞)
Y_3	ID del blocco della variabile base	intero [1,+∞)
Y_4	Data set della variabile base	testo
Y_5	Nome della variabile base	testo
Y_6	Chiave del data set base	testo
Y_7	Modalità della variabile base	intero (-∞,+∞)
Y_8	Minimo della variabile base	numerico (-∞,+∞)
Y_9	Massimo della variabile base	numerico (-∞,+∞)

La struttura di questo secondo file è del tutto simile a quella del file precedente e quindi tale che ogni riga fornisca le informazioni necessarie per la costruzione di una modalità della variabile ausiliaria. In particolare, tutte le righe del file ausiliario concernenti la stessa variabile ausiliaria e aventi variabili base con lo stesso ID sono in relazione logica di tipo *OR*, mentre tutti i record del file ausiliario concernenti la stessa variabile ausiliaria e aventi variabili base con ID differente sono in relazione logica di tipo *AND*.

Si possono quindi combinare più variabili base per costruire una modalità della variabile ausiliaria. Ad esempio, supponendo di dover costruire una modalità *a* della variabile ausiliaria *VAUX*1 sfruttando la variabile base qualitativa *VBASE*1 e la variabile base quantitativa *VBASE*2, entrambe contenute in un *data set* diverso da quello che si sta trattando, i campi y_4 e y_6 sono rispettivamente valorizzati con il nome del *data set* contenente le variabili base e con la chiave di aggancio tra questo ed il *data set* nel quale si vuol costruire la variabile ausiliaria. Nel file di metadati si scrivono dunque due record, uno per ciascuna variabile base; nel primo record si ha y_5=*VBASE*1 e y_7=b, mentre nel secondo record si ha y_5=*VBASE*2 e y_8=c e y_9=d. Entrambi i record riportano y_1=*VAUX*1 e y_2=a. Questi due record danno origine a due condizioni logiche del tipo "*if … then …*" e la colonna y_3=*ID* serve per distinguere se le due condizioni logiche devono essere combinate in *OR* oppure in *AND*.

Esempio 1:

y_1	y_2	y_3		y4	y_5	y_6	y_7	y_8	y_9
VAUX1	a	1	DSET1	VBASE1	KEYDSET1		b		
VAUX1	a	1	DSET2	VBASE2	KEYDSET2			c	d

-> if (VBASE1=b) ***OR*** *(c≤VBASE2≤d) then VAUX1=a*

Esempio 2:

y_1	y_2	y_3		y4	y_5	y_6	y_7	y_8	y_9
VAUX1	a	1	DSET1	VBASE1	KEYDSET1		b		
VAUX1	a	2	DSET2	VBASE2	KEYDSET2			c	d

-> if (VBASE1=b) ***AND*** *(c≤VBASE2≤d) then VAUX1=a*

In generale, in fase di esecuzione del *software*, le variabili ausiliarie vengono inserite nel *data set* delle interviste nelle posizioni indicate nel *routing file*, assicurando piena coerenza tra il grafo DAG e il *data set* dei microdati. A questo punto è possibile costruire la matrice di incidenza **I**, la matrice di adiacenza **A** e la matrice **D** dei gradi di ciascun vertice del grafo. Come già detto, queste matrici consentono l'applicazione di diversi algoritmi noti nell'ambito della Teoria dei Grafi. Per il sistema di generazione automatica di regole, la matrice di interesse è quella di incidenza.

4.4 Il metodo di correzione probabilistica adottato in EU-SILC

L'approccio della correzione probabilistica richiede la definizione dell'insieme di regole (*edit*) che sono necessarie per individuare un dato errore senza la necessità di enunciare tutte le azioni da intraprendere per correggere tale errore. La correzione dell'errore viene effettuata mediante l'utilizzo di opportuni algoritmi. Il *software* largamente usato in Istat per compiere tale tipo di correzione è SCIA (Sistema Controllo ed Imputazione Automatici; Istat, 2004) che implementa la metodologia *Fellegi-Holt* (Fellegi e Holt, 1976).

Questa metodologia prevede la costruzione di regole di incompatibilità dette anche *edit in forma normale*. Un *edit* è costituito dalla congiunzione di due o più condizioni sui valori che assumono le variabili del record del *data set* di microdati in esame. L'*edit* si attiva quando si verificano simultaneamente tutte le condizioni di incompatibilità che costituiscono l'*edit* stesso.

Tutti gli *edit* scritti dall'esperto prima di effettuare la correzione sono detti *edit espliciti* e sono necessari e sufficienti per l'individuazione dell'errore, ma non bastano per garantire una corretta imputazione dei valori delle variabili. Infatti, il principio di ottimalità impone che a seguito della modifica di certi valori di alcune variabili non si introducano nuovi errori, mentre il principio di minimalità impone che il numero delle variabili modificate sia il più basso possibile e non vi siano regole ridondanti. Nella individuazione e correzione dell'errore occorre considerare anche gli *edit impliciti*, cioè tutte quelle regole che si ricavano a partire dall'insieme degli *edit* espliciti mediante un processo di *derivazione* che porta alla definizione dell'*insieme minimo e completo* degli *edit*.

Sostanzialmente, il metodo *Fellegi-Holt* si concentra proprio sulla derivazione dell'insieme minimo e completo che matematicamente è un problema di ottimizzazione detto *copertura di insiemi (set covering).* Come noto in letteratura, tale problema rientra nella classe di complessità *NP-completo Hard* che vuol dire che non è possibile avere una stima a priori dello spazio di memoria e del tempo di esecuzione necessari per la derivazione dell'insieme minimo e completo. Possono verificarsi situazioni di blocco o di instabilità del *software* che comprometterebbero il buon esito della correzione probabilistica.

4.5 La generazione automatica degli *edit*

Data la complessità dei modelli di rilevazione dell'indagine EU-SILC, l'adozione efficiente dell'approccio di correzione brevemente riportato sopra impone la suddivisione del questionario in "sezioni" o "partizioni" di dimensioni opportune che dia luogo ad insiemi di *edit* espliciti piccoli quanto basta per non rendere impossibile la derivazione dei loro insiemi minimi e completi. Queste sezioni comportano l'individuazione di "passi" in cui suddividere la correzione degli errori delle variabili qualitative di EU-SILC. Un siffatto compito è di certo fattibile anche manualmente ma con costi più elevati in termini di tempo e di risorse umane.

Inoltre, qualora si debba modificare più volte la suddivisione del questionario in sezioni per migliorare la qualità della correzione o perché i passi individuati non conducono alla derivazione dei rispettivi insiemi minimi e completi, la generazione automatica di tutti gli *edit* espliciti di ogni passo di correzione risulta essere una strategia efficace. È stato dunque realizzato un *Sistema Automatico di Generazione Edit* (SAGE) che, a partire dai file di metadati di cui al paragrafo 4.3, costruisca la matrice di incidenza $\mathbf{I}=\{i_{ij}\}$ del grafo DAG (paragrafo 4.2) che viene utilizzata dal sistema per la generazione degli *edit* espliciti.

Si ricorda che la matrice di incidenza **I** ha le variabili in riga e tutte le modalità in colonna secondo l'ordine delle variabili. Ogni colonna della matrice **I** ha sempre solo due elementi non nulli in corrispondenza delle estremità dell'arco individuato da quella colonna[3]. Ad esempio, dalla variabile iniziale v_A *(*+1*)* si "salta" alla variabile finale v_X *(*-1*)* se la variabile iniziale assume la modalità i_{AX}, rappresentata dalla rispettiva colonna della matrice di incidenza. Le regole di incompatibilità che si ricavano sono del tipo: "*se* $v_A=i_{AX}$ *allora è un errore che* v_X *sia nullo*" e anche tutte le incompatibilità del tipo "*se* $v_A=i_{AX}$ *allora è un errore che* v_B, v_C, …, v_W *siano valorizzate*". Scorrendo, colonna per colonna, tutta la matrice si ottiene l'insieme di tutti gli *edit* espliciti che sono ricavabili proprio dalla struttura del questionario.

Grazie alla struttura descritta nel *routing file* e dunque la matrice di incidenza del questionario/grafo, si ottengono il file *VARDOM*[4] ed il file *REGOLE* necessari per effettuare la correzione probabilistica del passo k-esimo (generica sezione in cui si è suddiviso il questionario). È importante sottolineare che gli *edit* formali per la derivazione dell'insieme minimo e completo si ricavano sia dal file *REGOLE* che dal file *VARDOM*.

Il vantaggio di aver formalizzato il questionario come un grafo DAG è che, qualora sia necessario suddividere il questionario in sezioni, può essere usato un algoritmo *greedy* di partizionamento. Il sistema SAGE utilizza un partizionamento che individua quelle sezioni

3 Tutti quei vertici corrispondenti a domande del questionario che necessitano di memorizzare le risposte in gruppi logici di variabili di tipo *OR* oppure *XOR* sono trasformati opportunamente in un unico *super-vertice*.

4 I file *VARDOM* e *REGOLE* costituiscono l'input della procedura SCIA. Nel file *VARDOM*, oltre ai valori ammissibili di ciascuna variabile, sono indicate quelle variabili che ammettono il valore nullo e quelle che devono essere necessariamente riempite. Questa è una conseguenza diretta dall'analisi dei percorsi del questionario.

di questionario che sono collegate tra loro per il tramite di un solo arco (*taglio minimo*). È importante sottolineare che tale algoritmo non realizza un "partizionamento ottimo" del questionario che porterebbe ad avere partizioni con ugual numero di variabili e di percorsi. L'adozione di un algoritmo più semplice quale quello implementato in SAGE costituisce di fatto un buon supporto alle decisioni.

4.6 Un esempio di generazione automatica degli *edit* formali (EU-SILC 2016)

Allo scopo di fornire una valutazione dell'utilità di un approccio automatico quale quello qui descritto, si vuole mostrare la generazione degli *edit* che scaturiscono dalla struttura dei modelli di rilevazione utilizzati per raccogliere le informazioni sull'intera famiglia (scheda famigliare, questionario famigliare e questionario individuale).

Gli *edit* generati automaticamente sono *formali*, ovvero servono per verificare che le variabili assumano valori sempre compatibili con i percorsi del questionario. È importante puntualizzare che esistono anche *edit sostanziali* che nascono da condizioni di compatibilità di tipo logico e, di solito, non hanno una relazione diretta con i percorsi del questionario. Ad esempio, si supponga che a certe domande del questionario possano rispondere solo i soggetti di sesso femminile (come l'assegno di maternità per la nascita di figli) oppure che l'attività economica dell'unità presso la quale l'intervistato lavora sia "Pubblica amministrazione" e il settore di lavoro indicato sia "privato"; questi *edit* non propriamente formali non costituiscono un percorso in senso stretto. Le potenzialità del questionario elettronico permettono di implementare dei filtri di controllo già in fase di intervista tali da sottoporre il quesito solo ai soggetti che soddisfano le caratteristiche desiderate o di implementare delle avvertenze per il rilevatore per evitare l'inserimento di informazioni incompatibili; tali regole devono comunque essere rispettate anche in sede di controllo e correzione dei dati (piano di *check*). Sebbene, come appena descritto, nella maggior parte dei casi sia possibile trasformare un *edit* sostanziale in uno formale, a tutt'oggi in EU-SILC si preferisce velocizzare la produzione del file delle regole scrivendo manualmente gli *edit* sostanziali di particolare complessità da aggiungere a quelli formali.

Inoltre, l'indagine prevede annualmente l'aggiunta di moduli *ad-hoc* che approfondiscono ciclicamente tematiche specifiche differenti; i quesiti che afferiscono a tali moduli fanno riferimento a variabili e a *edit* che non necessariamente sono legati formalmente al resto del questionario. In questi casi sarebbe non opportuno adottare la procedura automatica di partizionamento, in quanto il modulo potrebbe essere indipendente per costruzione dalle altre sezioni del questionario; in questo caso, utilizzando l'approccio automatico, tale modulo verosimilmente verrebbe inglobato in una delle partizioni già individuate dalla procedura, con il rischio di inficiare la derivazione dell'insieme completo relativo a quella partizione. Tuttavia avere a disposizione un metodo per suddividere il questionario in sezioni rimane una funzionalità fondamentale per tentare di ovviare al problema della non derivabilità per tutti quei casi che non richiedono scelte ragionate a priori.

Nel corso degli anni, l'approccio è stato utilizzato con profitto sia nella versione PAPI di EU-SILC che nella sua versione CAPI e mista CAPI-CATI. Di seguito, si riportano i risultati della generazione degli *edit* formali del questionario CAPI-CATI del 2016, mettendo a confronto la generazione ottenuta dai questionari non partizionati con la generazione dalle sezioni ottimali prodotta da SAGE.

Il *routing file* della scheda famigliare (SF) consta di 110 variabili di interesse (file *VAR-DOM*) e sviluppa 709 *edit* formali (file *REGOLE*). Il *routing file* del questionario famigliare (QF) consta di 248 variabili di interesse e sviluppa 585 *edit* formali. Infine, il *routing file* del questionario individuale (QI) consta di 449 variabili di interesse e sviluppa 1.401 *edit* formali. I grafi DAG dei questionari hanno le seguenti matrici di incidenza (*nodi* × *archi*): $\mathbf{I}_{SF}$ (110 × 953), $\mathbf{I}_{QF}$ (248 × 592) e $\mathbf{I}_{QI}$ (449 × 1568).

Il *software* realizzato (SAGE) implementa un algoritmo di partizionamento di grafi diretti aciclici per la determinazione di sezioni (partizioni) del questionario; si noti che nell'ipotesi limite in cui l'intervistato debba rispondere a tutti i quesiti, ovvero in assenza di domande filtro, il percorso dalla sorgente al pozzo è univoco con un numero minimo di due variabili per partizione legate sequenzialmente l'una all'altra. Ovviamente nella pratica si vengono a determinare situazioni ben diverse da quella limite appena descritta: in Tavola 4.3 si riportano i risultati del partizionamento automatico dei tre modelli di rilevazione impiegati nell'indagine EU-SILC, ottenuti applicando l'algoritmo di partizionamento implementato nel *software* (in particolare, il numero di partizioni ottenute e i momenti principali delle distribuzioni delle variabili e dei percorsi possibili nelle varie partizioni).

Tavola 4.3 - Partizionamento automatico dei modelli di rilevazione di EU-SILC. Anno 2016

MODELLI DI RILEVAZIONE	STATISTICHE DI SINTESI	Partizionamento automatico	
		Variabili	Percorsi
Scheda famigliare			
	Valore medio	55,0	8.456,0
	Deviazione standard	10,6	11.675,8
	Valore minimo	48	200
	Valore massimo	63	16.712
	Numero di partizioni		2
Questionario famigliare			
	Valore medio	36,3	591,4
	Deviazione standard	20,7	1.353,4
	Valore minimo	15	24
	Valore massimo	63	3.660
	Numero di partizioni		7
Questionario individuale			
	Valore medio	35,5	822,9
	Deviazione standard	26,7	2.528,1
	Valore minimo	14	99
	Valore massimo	99	9.232
	Numero di partizioni		13

Fonte: Istat, Indagine sul reddito e le condizioni di vita delle famiglie

La scheda famigliare si suddivide in 2 partizioni, con un numero medio di 55 variabili per sezione; poiché l'algoritmo adoperato in SAGE "taglia" non appena la condizione di *taglio minimo* si verifica, il numero delle variabili nelle due sezioni passa da 48 a 63 con un consistente aumento dei percorsi che si sviluppano (si passa da 200 a 16.712 percorsi). Per dare un'idea della complessità del questionario, si riporta la stima del numero complessivo di percorsi della scheda famigliare che è circa $3{,}34 \times 10^6$ percorsi possibili. Il questionario famigliare è invece suddivisibile in 7 partizioni e sviluppa circa $5{,}04 \times 10^{14}$ percorsi. Il questionario individuale è suddivisibile in 13 partizioni e sviluppa circa $1{,}27 \times 10^{27}$ percorsi.

Dai risultati ottenuti, si evince che il partizionamento fornisce indicazioni utili sulla complessità dei questionari. Infatti, il numero di percorsi possibili del questionario è strettamente legato alle dimensioni dell'insieme minimo e completo e suddividere il questio-

nario in sezioni porta ad effettuare una correzione in k passi che coinvolgono ciascuno un numero minore di variabili piuttosto che in uno solo talmente complesso che potrebbe non consentire la derivazione dell'insieme minimo e completo. Ovviamente, non vi è alcuna garanzia che uno o più sezioni portino ad un set di regole tali da poter derivare il loro insieme minimo e completo; in tal caso, occorre aumentare la dimensione del taglio tra una sezione e l'altra, cioè il numero di percorsi distinti che le collegano. Situazioni del genere fanno propendere per l'adozione di soluzioni di tipo ragionato (in modo manuale) per l'individuazione delle partizioni migliori.

4.7 *Edit* sostanziali e conseguenze sulle partizioni generate automaticamente

La fase di controllo e imputazione probabilistica degli errori delle variabili qualitative richiede una serie di valutazioni preliminari che guidino in maniera ragionata il processo. È necessario, infatti, integrare le regole formali di percorso con le regole sostanziali di relazioni logiche e longitudinali tra variabili, tra cui l'imposizione di controlli e vincoli di coerenza con le informazioni provenienti da fonte amministrativa e da edizioni di indagine precedenti. Inoltre, le variabili da trattare presentano per natura livelli di rilevanza differenti che impongono di individuare *a priori* una suddivisione gerarchica in passi di correzione, alcuni dei quali dipendenti l'uno dall'altro per la presenza di variabili comuni. Ad esempio, le variabili del prospetto redditi che identificano il profilo di reddito percepito (dipendente, autonomo, da pensione, altro) fungono da filtro per l'accesso alle sezioni di pertinenza; è ovvio quindi che tali sezioni siano gerarchicamente dipendenti dal prospetto redditi, il che si riflette nel dare precedenza alla correzione di quest'ultimo. Tutto ciò fa sì che il sistema di generazione degli *edit* fin qui illustrato non possa essere affidato ad una procedura esclusivamente automatica, ma vada riadattato e affrontato in una certa misura anche per via manuale in modo ragionato.

La strategia di partizionamento adottata per EU-SILC consiste quindi nella ricerca in prima battuta di una suddivisione in passi distinti e indipendenti tra loro, ciascuno con un numero ragionevolmente basso di regole esplicite in modo che sia possibile derivare un insieme completo di regole per ciascuno di essi. Naturalmente la suddivisione sarà tanto migliore quanto minore è il numero di variabili comuni: il risultato ottimale è quello in cui i vari sottoinsiemi di regole generati risultino completamente disgiunti, e in tal caso l'ordine di esecuzione è ininfluente (Barcaroli *et al.*, 1999). Nel caso complesso del questionario EU-SILC questo non avviene in quanto alcuni passi contengono variabili comuni che li vincolano ad essere dipendenti l'uno dall'altro: in questo caso è necessario definire una sequenza ben precisa nell'ordine di esecuzione dei passi a seconda dei diversi livelli di importanza delle variabili coinvolte e delle relazioni tra loro esistenti; durante l'esecuzione di uno dei processi di elaborazione sarà necessario tenere fisse, ovvero non più suscettibili di cambiamenti, tutte le variabili imputate dai processi precedentemente eseguiti, considerate di filtro per quelle che da esse dipendono.

Formalmente, il partizionamento impone una gerarchia tra i passi individuati (coincidenti con le sezioni o partizioni del grafo). L'arco che collega due partizioni $\mathcal{A}$ e $\mathcal{B}$ fa sì che due variabili, ciascuna delle quali appartenente ad una partizione diversa, siano *adiacenti* (estremità dell'arco); occorre quindi che la variabile della partizione $\mathcal{A}$ che incide sull'arco suddetto, una volta corretti i valori delle variabili di $\mathcal{A}$ con SCIA, debba essere inglobata,

insieme all'*edit* che l'arco rappresenta, nella partizione $\mathcal{B}$ assumendo *fissità massima*[5] e poi procedere alla correzione della partizione $\mathcal{B}$.

Ciò è ancor più evidente se si introducono manualmente gli *edit* sostanziali, i quali potrebbero unire nuovamente le partizioni individuate. Le sezioni istruzione e lavoro, ad esempio, sono legate dalla regola logica "dichiara di frequentare un corso finanziato dall'azienda in cui lavora ma risulta non avere un lavoro". In generale, l'introduzione di questa tipologia di *edit* richiede una serie di valutazioni che, nell'intento di utilizzare l'approccio automatico, possono rendere oneroso il processo di correzione richiedendo un elevato numero di variabili ausiliarie per la trasformazione degli *edit* sostanziali in *edit* formali. Si preferisce quindi integrare le regole formali con le regole sostanziali aggiungendole manualmente al file *REGOLE*.

Inoltre, il fatto che le variabili da trattare presentino, per natura, livelli di rilevanza differenti, imponendo priorità diverse tra i passi di correzione, dà luogo a degli ulteriori *edit* sostanziali che tengano conto della fissità delle variabili; occorre valutare di volta in volta se convenga optare più per un approccio ragionato anziché automatico, il che non costituisce un limite dell'approccio proposto bensì una massimizzazione della sinergia di entrambi i criteri.

Dunque, la strategia di partizionamento adottata finora per EU-SILC consiste nella ricerca, in primo luogo, di una suddivisione automatica in passi distinti ottenuti considerando solo *edit* formali, per poi aggiungere manualmente gli *edit* sostanziali valutando quali partizioni sia più opportuno eventualmente unire o ulteriormente suddividere. Naturalmente, si cerca di bilanciare il numero di variabili e il numero di *edit* espliciti nel tentativo di ottenere dei passi di correzione per i quali non sia oneroso derivare l'insieme minimo e completo.

In Tavola 4.4 sono riassunti, per ciascuno dei tre modelli di rilevazione, il numero di partizioni, il numero di variabili e il numero di *edit* formali ottenuti con i due approcci, automatico e manuale, per l'individuazione dei passi di correzione dei dati di EU-SILC relativi all'anno di indagine 2016.

Tavola 4.4 - Confronto tra partizionamento automatico e manuale dei modelli di rilevazione di EU-SILC. Anno 2016

MODELLI DI RILEVAZIONE	STATISTICHE DI SINTESI	Partizionamento automatico		Partizionamento manuale		
		Variabili	*Edit* formali	Variabili	*Edit* formali	*Edit* sostanziali
Scheda famigliare						
	Valore medio	55,0	354,5	67,0	687,0	33,0
	Deviazione standard	10,6	427,8	-	-	-
	Valore minimo	48	52	67	687	33
	Valore massimo	63	657	67	687	33
	Numero di partizioni		2			1
Questionario famigliare						
	Valore medio	36,3	83,6	116,0	292,5	16,0
	Deviazione standard	20,7	75,0	113,1	229,8	14,1
	Valore minimo	15	15	36	130	6
	Valore massimo	63	243	196	455	26
	Numero di partizioni		7			2
Questionario individuale						
	Valore medio	35,5	104,8	46,9	115,9	35,2
	Deviazione standard	26,7	109,4	21,7	71,3	35,8
	Valore minimo	14	14	21	16	1
	Valore massimo	99	349	88	268	131
	Numero di partizioni		13			11

Fonte: Istat, Indagine sul reddito e le condizioni di vita delle famiglie

5 Il valore della fissità viene indicato nel *VARDOM* e quando è pari al valore massimo previsto in SCIA (i.e. 9) significa che la variabile deve mantenere inalterato il proprio valore durante la correzione probabilistica.

Si può notare immediatamente che la suddivisione dei questionari effettuata dagli esperti dell'indagine sulla base di considerazioni di importanza delle variabili e di regole sostanziali produce un minor numero di partizioni, ovvero predilige partizioni più grandi. È da aggiungere che la derivazione dell'insieme minimo e completo è stata possibile per tutti i passi individuati manualmente senza bisogno di alcuna modifica. È evidente che la procedura che individua automaticamente le partizioni tende a realizzare passi di correzione più piccoli quindi meno onerosi computazionalmente, ma è pur vero che questi sono stati ottenuti in assenza di *edit* sostanziali, che implicano una gerarchia ragionata dei passi di correzione più vincolante rispetto a quella che si avrebbe con i soli *edit* formali. È importante sottolineare che i passi di correzione ottenuti automaticamente, non tenendo conto in questo esempio degli *edit* sostanziali che è possibile definire solo manualmente, non portano a un livello soddisfacente di correzione dei dati; questo impone l'inserimento manuale di tale tipologia di *edit*, con conseguente riduzione del numero di passi.

La Tavola 4.4 mostra inoltre che, secondo i due approcci, si distribuiscono in maniera differente sia il numero di variabili sia il numero di *edit* formali. Infatti, l'approccio automatico produce una suddivisione in sezioni composte da un numero inferiore di variabili e di edit. Tuttavia, la fase di derivazione per i passi di correzione ottenuti manualmente è stata completata con successo così come per quelli ottenuti automaticamente. La distanza tra i due approcci è data quindi dalla presenza degli *edit* sostanziali: la "linearità" del questionario famigliare produce meno partizioni nell'approccio manuale nonostante la presenza degli *edit* sostanziali, mentre nel caso del questionario individuale i due approcci risultano simili in termini di numero finale di partizioni.

Lavorando manualmente si tende a preferire meno partizioni, quindi più grandi, rispetto a quelle che si individuano algoritmicamente, in quanto oltre alla facilità di gestione, proprio perché sono in numero contenuto, rispondono meglio al principio di correzione basato sul minimo cambiamento alla base della metodologia *Fellegi-Holt*. Il numero di *edit* sostanziali che sono stati aggiunti manualmente ad ogni partizione è, in media, pari a 33 per la scheda famigliare, a 16 per il questionario famigliare (valore *min* = 6 e valore *max* = 28), mentre per il questionario individuale è pari a 35,18 (valore *min* = 1 e valore *max* = 131). È intuitivo pensare che, con un approccio automatico e al netto degli *edit* sostanziali, aumentando il numero di partizioni queste diventino più snelle.

La combinazione tra approccio automatico e manuale permette di sfruttarne al massimo i vantaggi, riducendo al contempo i limiti derivanti da ciascuno di essi. Il *trade-off* tra costi e benefici è legato al fatto che la procedura automatica potrebbe trasformare matematicamente buona parte degli *edit* sostanziali in formali; ma di fatto l'onere che comporta la preparazione dei metadati, nonché la complessità dell'individuazione del giusto "taglio" del grafo conseguente all'aggiunta delle regole sostanziali o di nuove sezioni nel questionario, a tutt'oggi è tale che si preferisce agire in modo controllato ovvero manuale. Di contro, stante la complessità del questionario, sarebbe impensabile provvedere all'individuazione e alla generazione di tutti gli *edit* in modo manuale, soprattutto quelli formali; da qui il grande vantaggio di disporre di una procedura che consenta di farlo automaticamente, massimizzandone così la tempestività e minimizzandone al contempo l'errore umano.

4.8 Conclusioni

Poter disporre di un *software* basato sull'approccio automatico di generazione di *edit* è uno strumento di indubbia utilità. La modellazione matematica utilizzata è solida e permette di andare ben oltre lo scopo della generazione automatica degli *edit* formali o del partizionamento del questionario. Tale approccio può essere applicato a qualsiasi tipo di questionario, a prescindere dal tipo di somministrazione dell'intervista (con tecnica PAPI, o CAPI, ecc.).

La disponibilità di un questionario elettronico presenta il vantaggio di poter inserire parte del piano di verifica già al suo interno, garantendo un buon livello qualitativo dei dati grezzi, minimizzando così l'impatto del successivo piano di *check* sul file dati e riducendo ulteriormente l'errore non campionario. Tuttavia, nel questionario elettronico di EU-SILC vengono usati solo (o principalmente) controlli che permettono soltanto di segnalare eventuali incompatibilità (controlli di tipo *soft*), che non necessariamente riescono ad essere risolte in modo immediato e definitivo direttamente in fase di acquisizione dati; inoltre riguardano principalmente le coerenze di percorso del questionario e prendono solo marginalmente in considerazione le incompatibilità che possono emergere tra le informazioni raccolte in punti differenti del questionario stesso; infine vi è anche la necessità di conciliare le informazioni raccolte tramite l'intervista con alcune altre provenienti da archivi amministrativi. La struttura del questionario EU-SILC, come già mostrato nei paragrafi precedenti, comporta numerose regole di compatibilità; inoltre, i cambiamenti che nel tempo possono interessare il questionario, come l'inserimento di nuovi quesiti, la modifica di intere sezioni o, più semplicemente, la modifica di modalità di risposta, danno luogo a variazioni nella definizione di alcune regole sia formali sia sostanziali.

Da qui la necessità di disporre di un sistema di generazione automatica degli *edit* che permetta di gestire efficacemente l'intero piano di *check*, indicando quelle partizioni del questionario che singolarmente non risentano dei limiti computazionali nella derivazione dell'insieme completo degli *edit*. Nonostante ciò, poiché la correzione delle variabili qualitative adottata in EU-SILC combina il ricorso alla procedura proposta con l'aggiunta di numerosi *edit* sostanziali, la ricerca automatica di partizioni del questionario può rappresentare una soluzione di partenza per l'approccio ragionato. Si è optato quindi per una suddivisione "ragionata" in sezioni cui corrispondono dei passi di correzione gerarchicamente legati tra loro. Tale operazione, compiuta manualmente, è comunque notevolmente agevolata dallo sfruttamento delle partizioni già individuate con l'approccio automatico. L'approccio ragionato, in definitiva, nonostante non sia in grado di garantire che tra una sezione e la successiva vi sia un solo arco di collegamento, condizione necessaria e sufficiente per il taglio tra le due partizioni secondo la suddivisione automatica, si dimostra efficace al fine di assicurare la massima coerenza possibile tra le informazioni rilevate tramite le interviste, quelle ricavate dalle fonti amministrative e quelle relative alle precedenti occasioni di indagine.

L'approccio proposto permette di ricavare la struttura matematica del questionario che è alla base non solo della procedura di generazione automatica degli *edit* (formali) qui presentata, ma anche di strumenti utili per investigare la complessità del questionario e per il controllo di compatibilità di un *data set* con la struttura del questionario a cui esso si riferisce.

5. FRI: UNA NUOVA PROCEDURA PER L'IMPUTAZIONE DA DONATORE[1]

5.1 Il controllo e la correzione dei dati

In questo capitolo è illustrato un approccio deterministico per l'imputazione da donatore utilizzata nel processo di trattamento dati di EU-SILC per l'imputazione totale dei record completamente privi di valori oppure per l'imputazione di un insieme predefinito di variabili (i cui valori sono mancanti o ritenuti inaffidabili). Tale approccio è usato laddove si ritiene opportuno prendere tutti i valori da un record donatore reputato più "vicino" al record da correggere (record ricevente), al fine di trasferire a quest'ultimo una sequenza di risposte ai diversi quesiti considerata plausibile, in quanto già realizzata. La procedura qui descritta implementa la metodologia alla base anche del software RIDA (Barcaroli *et al.*, 1999), conosciuto in Istat e adottato nel trattamento dati di EU-SILC nei primi anni di indagine. La nuova procedura per l'imputazione da donatore introdotta nel processo produttivo che verrà illustrata di seguito, presenta degli indubbi vantaggi: è maggiormente integrata con le restanti procedure di correzione dei dati, prevalentemente implementate nell'ambiente software SAS, non richiede pertanto l'installazione sul proprio computer di software aggiuntivo. A tale beneficio, si aggiunge la facilità di utilizzo. RIDA, d'altro canto realizzato anni orsono, non è facilmente adattabile oggi sia nella versione *batch* (eseguibile su piattaforme Windows con versione del sistema operativo che non va oltre la versione XP) che in quella SAS (programma CONCORD non più supportato da assistenza) o quella Java che richiede la presenza di software aggiuntivo, che deve essere frequentemente aggiornato per non interferire con le funzionalità del sistema operativo del computer. Nel Paragrafo 5.2 si riportano alcuni cenni metodologici sul tipo di metrica utilizzata per individuare il donatore "migliore" per ciascun record da imputare. Il Paragrafo 5.3 spiega come è articolato l'algoritmo che sta alla base del software realizzato per la correzione dei dati EU-SILC. Il Paragrafo 5.4 descrive la tipologia di output e dei report prodotti dalla procedura. Il Paragrafo 5.5 riporta, come esempio di applicazione, i risultati dell'imputazione effettuata sui dati della rilevazione condotta nell'anno 2012, ovvero l'anno in cui è stata introdotta la procedura qui presentata. Alcune considerazioni conclusive e possibili sviluppi dell'approccio proposto sono riportati nel Paragrafo 5.6.

5.2 Cenni metodologici

In diverse occasioni, durante il processo di correzione dei dati, può diventare necessario sostituire i valori di alcune variabili per risolvere alcune incompatibilità e, nei casi limite di dover sostituire un intero *data record*. Un metodo usato per risolvere questo tipo di situazioni è quello della imputazione totale dei record da donatore. È un metodo deterministico che separa i dati in due insiemi, donatori e riceventi, per poi individuare il miglior dona-

1 Il capitolo è stato redatto da Pierpaolo Massoli.

tore per ciascun ricevente cioè il donatore a "*distanza minima*" dal ricevente (donatore *ottimo*). Nella procedura qui presentata, il criterio utilizzato per individuare il donatore ottimo è quello di *massima similarità* tra due unità statistiche. Come noto, il concetto di similarità è meno forte di quello di distanza ma sicuramente più corretto laddove le due unità statistiche abbiano attributi di tipo qualitativo. Quindi, nel caso di variabili qualitative è più appropriato adottare il concetto di similarità mentre nel caso di variabili quantitative è più opportuno utilizzare il concetto di distanza (Euclidea, Manhattan, Mahalanobis, ecc.). Inoltre, nel caso delle variabili qualitative, il trattamento di variabili dicotomiche, politomiche e ordinali richiede diverse valutazioni della similarità. Nello scenario di EU-SILC, data la presenza di variabili qualitative e quantitative, si è scelto di misurare la vicinanza tra due unità statistiche valutando la loro similarità mediante il calcolo dell'*indice di Gower* (Gower, 1971). La similarità tra due vettori composti da *m* elementi di cui *l* di tipo qualitativo e *n* di tipo quantitativo indicati con

$$\mathbf{x} = \{x_1^{ql}, \ldots, x_l^{ql}, \ldots, x_1^{qn}, \ldots, x_n^{qn}\} \quad \text{e} \quad \mathbf{y} = \{y_1^{ql}, \ldots, y_l^{ql}, \ldots, y_1^{qn}, \ldots, y_n^{qn}\}$$

è definita come:

$$s(\mathbf{x},\mathbf{y}) = \frac{1}{m}\left\{\sum_{i=1}^{l} I\left(x_i^{ql}, y_i^{ql}\right) + \sum_{j=1}^{n}\left(1 - \frac{\left|x_j^{qn} - y_j^{qn}\right|}{Range_j^{qn}}\right)\right\}$$

La funzione indicatrice $I(x_i^{ql}, y_i^{ql})$ vale 1 se $x_i^{ql}=y_i^{ql}$ e 0 altrimenti mentre il $Range_j^{qn}$ è la differenza tra il valore massimo ed il valore minimo della *j-esima* variabile quantitativa. Il valore dell'indice varia tra 0 (similarità nulla) e 1 (similarità massima) e questo facilita la lettura dei risultati nell'analisi dell'efficacia della donazione. È di facile implementazione ed è sempre valido per misurare la similarità di due vettori, sia di variabili qualitative, sia di variabili quantitative.

5.3 La procedura di imputazione totale da donatore

Il calcolo della massima similarità tra due record è utilizzato in EU-SILC in tutti quei casi in cui occorre, ad esempio, recuperare delle informazioni per il completamento parziale o totale dell'intervista di un individuo facente parte del campione dell'indagine. Si tratta di un approccio deterministico in cui le informazioni da recuperare per correggere le anomalie di un *data record (ricevente)* sono scelte dal *data record* più simile che non presenta anomalie (*donatore*). Individuato il donatore, le anomalie del ricevente vengono corrette imputando i valori delle variabili prese dal donatore. L'imputazione è totale se l'intero *data record* ricevente è sostituito con i valori del record del donatore, altrimenti è *parziale*; la scelta dipende dalla situazione specifica in fase di correzione dei dati e comunque è ininfluente nella determinazione del donatore. Nel caso specifico qui trattato, il metodo di imputazione descritto riguarda la correzione di alcune informazioni incompatibili che provengono dalle interviste individuali. La fase di correzione è articolata in tre passi: (1) preparazione dei *data set* di donatori e di riceventi (record candidati all'imputazione), (2) individuazione del donatore ottimo per ciascun ricevente e (3) costruzione dei *data set* corretti (output della procedura). Il

software che implementa tale approccio è chiamato FRI (*Full Record Imputation*), è scritto interamente in SAS ed è suddiviso in tre programmi che vengono eseguiti sequenzialmente:

1. *DONATORI_RICEVENTI*; nei *data set* da correggere si separano i *data record* da correggere (*data set RICEVENTI*) da quelli utilizzabili per l'imputazione (*data set DONATORI*);
2. *FULL_RECORD_IMPUTATION*: per ogni ricevente si calcola l'indice di similarità tra il ricevente e tutti i potenziali donatori;
3. *CREA_FILES_FASE70*[2]: si ricostruiscono i *data set* secondo la loro struttura originaria con i valori corretti, cioè i valori presi dal donatore individuato e sostituiti a quelli non corretti del ricevente.

Per l'esecuzione della procedura, è essenziale che le variabili di *match* (v.*infra*) per la valutazione della similarità e le variabili di strato per circoscrivere i gruppi di donatori non abbiano valori mancanti; in caso contrario, il software scarta quei record con valori mancanti. Pertanto, è necessario che precedenti fasi di controllo e correzione garantiscano tale requisito. Il criterio di suddivisione tra donatori e riceventi dipende solo da alcune variabili ritenute sufficientemente caratterizzanti un individuo presente nell'indagine. Nello specifico, il donatore viene ricercato tra gli "strati" in cui si suddivide il *data set* dei donatori in base ai valori di variabili di strato scelte dall'utente. Sono di solito variabili di tipo qualitativo e, ovviamente, devono essere sempre valorizzate sia nel *data set* dei donatori sia in quello dei riceventi. Dato un ricevente, i valori delle variabili di strato devono coincidere con quelli rispettivi del donatore. Individuato lo strato, il donatore migliore è quello più simile al ricevente sulla base del valore dell'indice mostrato nel Paragrafo 5.2, calcolato usando delle variabili di *match* sempre scelte dall'utente. Le variabili di *match* possono essere sia qualitative sia quantitative. Nel caso in cui per un ricevente vi siano due o più donatori possibili, con lo stesso valore di similarità, se ne sceglie casualmente uno, con una densità di probabilità uniforme. Come già detto, individuato il donatore si può decidere se "donare" al ricevente tutte le variabili pertinenti all'intervista individuale o solo una parte. È importante sottolineare che, il secondo passo della procedura (calcolo della massima similarità ed individuazione del donatore) può essere ripetuto iterativamente più volte, modificando la lista delle variabili di strato e/o di *match*: ciò risulta utile qualora, in prima battuta, non siano stati individuati donatori in alcuni strati; la rimozione di alcune variabili per la costruzione degli strati può consentire così l'individuazione di un donatore per ogni ricevente. Il secondo passo quindi adotta lo stesso approccio implementato dal software RIDA. In quest'ultimo si possono scegliere diversi tipi di metrica a seconda della natura dei dati e delle relazioni tra le variabili, mentre in FRI si è scelto di adottare il metodo più appropriato per il tipo di variabili trattate nell'indagine EU-SILC, il quale si presta, tuttavia, all'impiego in un'ampia generalità di casi per i motivi già visti. I programmi sono degli *script* SAS corredati da commenti, quindi sono leggibili ed eventualmente modificabili. Nel tentativo di ridurre tempi e possibili fonti d'errore, non è richiesta la creazione di file di input in formati diversi dal formato SAS. Fanno eccezione solo due file in MS Excel contenenti (i) la lista delle variabili di strato e di *match* e (ii) la lista delle variabili del ricevente che non devono ricevere i valori del donatore, nell'ipotesi di donazione parziale di alcune variabili dell'intervista individuale.

2 Per meglio documentare il processo di correzione in EU-SILC la generica fase di lavorazione **fase**[*k*] è contraddistinta da un numero *k*=10,20,30,…,100 ed eventuali altri numeri interi compresi tra 10 e 100.

5.4 Gli output della procedura

Lo scopo primario della procedura FRI è quello di produrre i *data set* contenenti i dati corretti secondo la struttura che questi avevano prima di iniziare la procedura. In generale, ogni *data record* contiene un apposito attributo detto chiave identificativa. Quando l'individuazione di tutti i donatori è completata, dopo aver eseguito più volte il secondo passo descritto al Paragrafo 5.3 (cambiando i criteri di individuazione), la procedura individua ad ogni *i-esima* esecuzione l'insieme di record da donare tra quello dei donatori (*data set* di input *DONATORI*) e quindi genera il *data set* con le sole chiavi identificative dei record donatori affiancate a quelle dei record riceventi (*CHIAVI_DONATE*) e il *data set* dei riceventi per i quali non è stato possibile trovare un donatore sulla base delle variabili di strato scelte per la *i-esima* esecuzione (RIC_NO_DON). Quest'ultimo *data set* viene usato come input dell'esecuzione *(i+1)-esima* costituendo il *data set* dei riceventi cui bisogna trovare ancora un donatore. Il *data set* dei donatori rimane sempre lo stesso di partenza, mentre il numero dei donatori effettivamente utilizzati nella specifica iterazione cambia a seconda delle variabili scelte per stratificare la popolazione dei donatori. Per meglio analizzare i risultati e investigare l'efficacia dell'imputazione, la procedura genera anche una serie di report inerenti ciascuna iterazione del passo di ricerca dei record donatori. I report prodotti sono:

- *01_VERIFICA INTEGRITÀ STRATO:* report per controllare che tutte le variabili di strato non presentino valori mancanti (sia nel *data set* dei donatori che in quello dei riceventi);
- *02_DETTAGLIO_ANALISI_STRATI:* report del numero degli strati in comune tra donatori e riceventi, degli strati nel file dei donatori che non hanno riscontro nel *data set* dei riceventi e, viceversa, strati nel *data set* dei riceventi che non trovano corrispettivi nei donatori;
- *03_DETTAGLIO_RICEVENTI_SENZA_DONATORE:* report del numero di riceventi per strato senza alcun donatore perché non esiste uno strato corrispettivo nel *data set* dei donatori;
- *04_DETTAGLIO_DONATORI_EFFETTIVI:* report dei donatori per strato effettivamente utilizzati per l'imputazione;
- *05_DETTAGLIO_RICEVENTI_EFFETTIVI:* report sui riceventi per strato che hanno trovato un donatore;
- *06_DETTAGLIO_STRATI_NUMERO_INFERIORE_RICEVENTI:* report degli strati con numerosità di donatori inferiore al numero dei riceventi per i quali l'imputazione potrebbe essere critica;
- *07_REPORT_ANALISI_STRATI:* report sintetico;
- *08_VERIFICA INTEGRITÀ MATCH:* report per controllare che tutte le variabili di *match* non abbiano alcun valore mancante (sia nel *data set* dei donatori che quello dei riceventi);
- *09_ANALISI_VARIABILI_MATCH:* report sui momenti della distribuzione delle variabili di *match*;
- *10_DISTRIBUZIONE_CHIAVI_DONATE:* report sul numero di volte in cui è stato utilizzato uno stesso donatore, distinguendo il numero di volte in cui esso è stato utilizzato con *match* perfetto (distanza nulla, indice di similarità massima) e quelle in cui è stato utilizzato senza raggiungere un *match* perfetto;
- *11_ANALISI_DELLA_DISTANZA:* report sulla distribuzione della distanza quando questa è non nulla;
- *12_CHIAVI_DONATE:* tabella associativa delle chiavi identificative dei riceventi e delle chiavi identificative dei donatori.

5.5 Un'applicazione ai dati dell'indagine EU-SILC

Occorre premettere che nell'indagine EU-SILC ci sono due unità di rilevazione: le famiglie e gli individui di 16 anni o più. Anche in presenza di un'unità familiare rispondente, può aversi il caso di un'unità individuale non rispondente: questa casistica, individuata come '*partial unit non-response*' dalle linee guida Eurostat (European Commission, 2017, pg.54), può essere gestita o attraverso un intervento sui coefficienti di riporto all'universo individuali, oppure attraverso un'operazione di imputazione: "*The term 'partial unit non-response' is introduced to describe the situation where some but not all individual members of a household selected for the survey have been successfully enumerated. Two possible approaches of dealing with this problem are: (i) adjustment of sample weights of enumerated individuals in the household with the objective of compensating for members not enumerated; or (ii) construction of the required variables for each non-enumerated person in the household through imputation.*".

La scelta italiana di procedere attraverso un'imputazione, anziché tramite una rettifica dei pesi individuali, è stata determinata dalla volontà di agevolare l'analisi dei dati e di disporre di un solo coefficiente di riporto all'universo, identico per la famiglia e tutti gli individui di cui essa si compone, piuttosto che coefficienti di riporto differenziati da utilizzare a seconda del diverso livello individuale o familiare dell'informazione da analizzare. Inoltre la numerosità esigua dei casi da trattare è stato un altro elemento che ha favorito l'opzione dell'imputazione.

Con riferimento all'applicazione qui descritta, i dati per la costruzione dei *data set* dei donatori e dei riceventi sono stati presi dai tre *data set* contenenti le osservazioni ricavate dalla scheda generale, il questionario familiare e il questionario individuale. Sia i donatori sia i riceventi sono determinati in base all'*eleggibilità* di un individuo e all'esito della sua intervista. Un individuo di almeno 16 anni è eleggibile se presente in famiglia o temporaneamente assente al momento della rilevazione. L'esito dell'intervista è favorevole quando l'individuo risponde direttamente all'intervistatore o, se temporaneamente assente, un altro membro (eleggibile) della famiglia risponde al suo posto (intervista *proxy*). Nel caso di questa applicazione, sono potenziali donatori tutti gli individui eleggibili che collaborano all'intervista mentre sono candidati riceventi gli individui eleggibili che rifiutano di collaborare o gli individui per i quali l'intervista diretta o *proxy* è impossibile. Questi record riceventi sono dunque incompleti, in quanto per essi sono disponibili solo le informazioni individuali fornite dal rispondente familiare, e, per risolvere le incompatibilità cui essi danno luogo, si è deciso di prelevare le informazioni mancanti dai record donatori, in linea con una delle possibili opzioni offerte dalle indicazioni metodologiche Eurostat summenzionate. Per completare la fase di individuazione dei donatori sono state necessarie due iterazioni della procedura. Il *data set* dei riceventi nella seconda iterazione è il *data set RIC_NO_DON* dell'iterazione precedente. Le condizioni per individuare i donatori nelle due iterazioni sono ovviamente diverse poiché con le variabili di strato e di *match* scelte per la prima non è stato possibile trovare un donatore per tutti i riceventi ed è stato necessario ridurre le condizioni suddette per completare la ricerca. Le variabili di strato e di *match* della prima esecuzione della procedura sono in Tavola 5.1 mentre quelle della seconda sono in Tavola 5.2.

Tavola 5.1 - Variabili di strato e di match della prima iterazione

Nome variabile	Ruolo	Tipo	Descrizione
STRAT15	strato	qualitativa	Numero di componenti della famiglia
RIP	strato	qualitativa	Ripartizione geografica
CLETA	strato	qualitativa	Età dell'individuo (in classi)
SEX	strato	qualitativa	Sesso dell'individuo
CITTADX	strato	qualitativa	Cittadinanza dell'individuo
ITALIA	strato	qualitativa	Individuo nato in Italia ?
FINEMESE	strato	qualitativa	La famiglia arriva alla fine del mese?
STACIV	strato	qualitativa	Stato civile dell'individuo
REDNET_E	match	quantitativa	Reddito individuale (netto)
GODAB_B	strato	qualitativa	Titolo di godimento dell'abitazione

Fonte: Elaborazioni su dati Istat

Tavola 5.2 - Variabili di strato e di *match* della seconda iterazione

STRAT15	strato	qualitativa	Numero di componenti della famiglia
RIP	strato	qualitativa	Ripartizione geografica
ETA	match	qualitativa	Età dell'individuo (variabile continua)
SEX	strato	qualitativa	Sesso dell'individuo
CITTADX	match	qualitativa	Cittadinanza dell'individuo
ITALIA	strato	qualitativa	Individuo nato in Italia ?
FINEMESE	match	qualitativa	La famiglia arriva alla fine del mese ?
STACIV	match	qualitativa	Stato civile dell'individuo
REDNET_E	match	quantitativa	Reddito individuale (netto)
GODAB_B	match	qualitativa	Titolo di godimento dell'abitazione
GODAB_B	strato	qualitativa	Titolo di godimento dell'abitazione

Fonte: Elaborazioni su dati Istat

I report di riepilogo, delle distribuzioni delle distanze e del numero di volte in cui è stato lo stesso donatore è stato accoppiato ad un ricevente delle due iterazioni sono riportate di seguito.

Tavola 5.3 - Riepilogo della prima iterazione

Indicatore	Valore
Numero complessivo di potenziali donatori	39261
Numero complessivo di potenziali riceventi	130
Numero effettivo di strati utilizzati per la donazione	101
Numero effettivo di donatori	3114
Numero medio di donatori utilizzati per strato	30,83
Numero massimo di donatori utilizzati per strato	178
Numero minimo di donatori utilizzati per strato	1
Numero di strati con un solo donatore	14
Numero di strati (in comune) con un numero di donatori = numero di riceventi	13
Numero di strati (in comune) con un numero di donatori < numero di riceventi	1
Numero effettivo di riceventi	116
Numero di riceventi senza un donatore	14

Fonte: Elaborazioni su dati Istat

Tavola 5.4 - Distribuzione della distanza tra ricevente e donatore nella prima iterazione

Similarità	Frequenza	Percentuale
similarità = 1	64	55,17
0<similarità< 1	52	44,83
Totale	116	100,00

Fonte: Elaborazioni su dati Istat

Tavola 5.5 - Numero di volte che lo stesso donatore è accoppiato ad un ricevente nella prima iterazione

Numero di volte	Frequenza	Percentuale
1	106	91,38
2	10	8,62
Totale	116	100,00

Fonte: Elaborazioni su dati Istat

Come evidenziato dai report, il numero di riceventi che sono soggetti alla donazione sono un numero esiguo rispetto al numero di record che possono essere presi come donatori (circa lo 0,33% del totale) e, in base alle variabili di strato indicate in Tavola 5.3 si individuano ben 101 strati anche se i donatori non sono equamente distribuiti tra loro. Per quei 116 riceventi cui è stato trovato un donatore il valore delle variabili di *match* in ciascun strato è tale che la similarità è massima nel 55,17% dei casi (*match* perfetto all'interno dello strato trovato) (Tavola 5.4). La donazione è avvenuta nella maggior parte dei casi (91,38%) utilizzando donatori sempre diversi (Tavola 5.5). La prima esecuzione della procedura non è sufficiente a completare la fase di correzione ma occorre ripetere l'esecuzione del passo 2 della procedura per ricercare "nuovi" donatori per i 14 riceventi per i quali la prima esecuzione non ha trovato alcun donatore. Le condizioni di ricerca dei donatori per i riceventi della nuova esecuzione del secondo passo della procedura sono in Tavola 5.2. I risultati sono nelle Tavole 5.6, 5.7 e 5.8.

Tavola 5.6 - Variabili di strato e di *match* della seconda iterazione

Indicatore	Valore
Numero complessivo di potenziali donatori	39261
Numero complessivo di potenziali riceventi	14
Numero effettivo di strati utilizzati per la donazione	10
Numero effettivo di donatori	4841
Numero medio di donatori utilizzati per strato	48,41
Numero massimo di donatori utilizzati per strato	1348
Numero minimo di donatori utilizzati per strato	64
Numero di strati con un solo donatore	0
Numero di strati (in comune) con un numero di donatori = numero di riceventi	0
Numero di strati (in comune) con un numero di donatori < numero di riceventi	0
Numero effettivo di riceventi	14
Numero di riceventi senza un donatore	0

Fonte: Elaborazioni su dati Istat

Tavola 5.7 - Distribuzione della distanza tra ricevente e donatore nella seconda iterazione

Similarità	Frequenza	Percentuale
0< similarità <1	14	100
Totale	14	100

Fonte: Elaborazioni su dati Istat

Tavola 5.8 - Numero di volte che lo stesso donatore è accoppiato ad un ricevente nella seconda iterazione

Numero di volte	Frequenza	Percentuale
	14	100
	14	100

Fonte: Elaborazioni su dati Istat

La seconda esecuzione del secondo passo della procedura completa l'individuazione di un donatore anche per ciascuno dei 130 riceventi iniziali poiché sono state ridotte le condizioni di strato e di *match* da rispettare, come mostrato in Tavola 5.2. Infine, il terzo passo della procedura proposta è specifico dell'indagine e, sostanzialmente, è necessario per prelevare dai donatori trovati solo le variabili che si vogliono correggere con il metodo proposto. Gli output di tale passo sono i *data set* dei microdati aventi la stessa struttura dei *data set* originale, ma con alcuni valori "donati".

5.6 Conclusioni

L'approccio proposto costituisce uno strumento efficace per la correzione dei dati mediante l'imputazione dei record mancanti o dei valori errati che dunque vengono presi dai record che non presentano errori. Sebbene il criterio che stabilisce quali siano i record corretti e che divide i *data set* dei microdati in donatori e riceventi sia specifico dell'indagine, l'approccio rimane generale e facilmente adattabile alle varie esigenze. La misura della similarità mediante l'indice adottato considera correttamente la presenza di variabili sia di tipo qualitativo sia di tipo quantitativo nei *data record* da trattare. Un ulteriore punto di forza di tale procedura è la disponibilità di un codice che non richiede compilazioni, installazioni di software aggiuntivo ed è scritto in un linguaggio largamente diffuso in Istat. Quest'ultimo aspetto rende la procedura FRI particolarmente vantaggiosa per coloro che trattano i dati in SAS System in quanto garantisce miglioramenti di produttività evitando "deviazioni di percorso" ovvero scrittura e lettura di dati in formato diverso da quelli accettati da SAS, soprattutto se la correzione dei dati richiede diverse re-iterazioni del secondo passo della procedura per l'individuazione dei donatori migliori per ciascun ricevente. Un possibile sviluppo della procedura FRI potrebbe essere la considerazione di altre metriche oltre all'indice di similarità di Gower, aprendo quindi alla possibilità di trattare anche dati testuali. Sarebbe inoltre utile disporre di una reportistica anche di tipo grafico che aiuti a visualizzare l'impatto della correzione.

6. LA STIMA DELLE TASSE E DEI CONTRIBUTI SOCIALI IN EU-SILC[1]

6.1 Introduzione

Nell'indagine EU-SILC che, come noto, utilizza congiuntamente i dati campionari e gli archivi amministrativi per la stima dei redditi al netto e al lordo dell'imposizione fiscale, l'applicazione del modello di micro-simulazione SM2-EU-SILC costituisce parte integrante della metodologia utilizzata.

Il modello fornisce, innanzitutto, la stima dei contributi sociali posti a carico dei lavoratori e dei datori di lavoro sull'intero campione di indagine.

Inoltre, stima le imposte sia per i componenti delle famiglie non inclusi nelle liste anagrafiche del campione, ma presenti al momento dell'intervista, sia per gli altri individui non agganciati agli archivi (ad esempio per errori nella generazione del codice fiscale)[2].

Infine, il modello stima le imposte anche per gli individui presenti negli archivi fiscali, ma le cui informazioni sulle imposte non sono utilizzate nel processo produttivo perché, ad esempio, non coerenti o conformi alle procedure di controllo e correzione utilizzate.

La metodologia di stima dei redditi lordi EU-SILC, basata su una validazione incrociata delle stime micro-simulate e dei dati delle dichiarazioni fiscali, consente di migliorare la qualità dei dati finali, come risulta anche dal confronto con i benchmark di Contabilità Nazionale. Tuttavia, quasi fin dall'inizio della produzione dei redditi lordi dell'indagine, per i contributi sociali è stato deciso di utilizzare esclusivamente la stima da modello a causa di una parziale copertura delle fonti amministrative attualmente disponibili in Istituto, in particolare per la contribuzione degli autonomi ma anche dei lavoratori dipendenti e parasubordinati.

Questo capitolo risulta quindi strutturato come segue: il secondo paragrafo presenta alcuni aspetti rilevanti della metodologia di stima del carico fiscale e contributivo delle famiglie nell'indagine EU-SILC. Il paragrafo 3 è dedicato alla procedura di stima della contribuzione obbligatoria prevista dall'ordinamento per il conseguimento delle prestazioni previdenziali e assistenziali, con la parametrizzazione delle aliquote contributive a carico dei lavoratori dipendenti e autonomi utilizzata nel modello SM2-EU-SILC. Il paragrafo 4 presenta un focus sulle imposte micro-simulate, con un'analisi della copertura delle informazioni derivanti da archivio fiscale e un confronto tra le stime finali dell'imposta e le stime da modello per una valutazione più approfondita del metodo utilizzato per la produzione dei redditi lordi EU-SILC.

6.2 Alcune specificità della metodologia di stima dell'imposizione fiscale e contributiva

I modelli di micro-simulazione, che consentono di stimare il carico fiscale e contributivo di individui e famiglie, rappresentano la tecnica comunemente utilizzata per la conversio-

1 Il capitolo è stato curato da Gabriella Donatiello.

2 L'aggancio agli archivi amministrativi viene effettuato tramite il codice fiscale, il quale viene fornito dalle liste anagrafiche comunali (LAC) per gli individui presenti anche nella famiglia anagrafica, oppure generato sulla base delle informazioni rilevate su cognome, nome, sesso, data e luogo di nascita per i componenti non anagrafici della famiglia.

ne netto/lordo dei redditi di un'indagine campionaria. All'esordio dell'indagine EU-SILC, la Commissione europea ha adottato il modello di micro-simulazione dell'Università di Siena SM2 come procedura raccomandata per la costruzione delle variabili lorde dell'indagine.

L'Istat ha tuttavia sperimentato una metodologia più complessa per stimare il carico delle imposte e dei contributi delle famiglie, adoperando congiuntamente il modello SM2 dell'Università di Siena e i dati relativi all'integrazione delle fonti campionarie e amministrative, attraverso tecniche di *record linkage*[3]. Sulla base di una metodologia consolidata, il modello di micro-simulazione stima le imposte per tutti i percettori di reddito, tali stime sono poi comparate con i dati fiscali e, in presenza di dati amministrativi ritenuti coerenti, sono sostituite con le imposte derivanti dalle dichiarazioni fiscali[4]. Le variabili lorde dell'indagine sono quindi costruite come somma dei redditi netti e delle imposte e ritenute di fonte amministrativa, se disponibili e coerenti, oppure come somma dei redditi netti e delle imposte micro-simulate[5]. I contributi sociali a carico dei lavoratori e dei datori di lavoro sono, invece, stimati unicamente dal modello, a causa di una copertura solo parziale degli archivi amministrativi attualmente disponibili in Istituto.

La disponibilità di stime micro-simulate e dei dati derivanti dalle dichiarazioni fiscali permette sostanzialmente di comparare e validare i risultati, migliorando la qualità delle stime finali dell'indagine.

Il confronto, in termini di composizione percentuale, delle stime finali dei redditi lordi EU-SILC con i dati della Contabilità Nazionale[6] (Tavola 6.1) mostra, per i redditi del 2014, una sostanziale coerenza delle stime, spiegata da alcuni fattori che concorrono a tale risultato. La Contabilità Nazionale utilizza tra le diverse fonti anche gli archivi amministrativi che sono integrati nel processo produttivo dell'indagine EU-SILC; in quest'ultima, l'utilizzo congiunto di redditi fiscali e redditi dichiarati all'indagine produce redditi mediamente più elevati di quelli amministrativi. Tuttavia in EU-SILC, come in tutte le indagini sui redditi, i redditi da capitale, e in particolare i redditi da attività finanziarie soggetti ad una ritenuta alla fonte, risultano sottostimati. Va considerato, infine, che le stime di Contabilità Nazionale includono l'economia sommersa, che parzialmente viene rilevata anche nell'indagine EU-SILC dove i rispondenti possono dichiarare redditi frutto di comportamenti di elusione e/o evasione fiscale e quindi per definizione non inclusi nei dati amministrativi.

Tavola 6.1 - Disaggregazione dei redditi lordi dell'indagine EU-SILC e confronto con i dati di Contabilità Nazionale (CN). Anno 2014 (valori in euro e percentuali)

	EU-SILC		CN	DIFFERENZA %
	(PRO CAPITE)			(EU-SILC-CN)
Reddito al lordo delle imposte e dei contributi sociali (a)	21.758	100,0	16,6	1,2
Contributi sociali	3.873	17,8	11,4	0,9
- Datori di lavoro	2.676	12,3	2,9	0,3
- Lavoratori dipendenti	708	3,3	2,3	0,0
- Lavoratori autonomi	489	2,2	82,7	-0,5
Reddito al lordo delle imposte	17.884	82,2	13,9	0,1
Imposte sul reddito e sulle attività finanziarie	3.052	14,0	68,8	-0,6
Reddito netto	14.833	68,2	100.0	100.0

(a) Include affitti imputati, autoconsumo, fringe benefits e contributi sociali a carico del datore di lavoro.

3 Per una descrizione dettagliata della metodologia di stima dei redditi lordi dell'indagine EU-SILC si veda Donatiello G., 2011.

4 Si utilizzano le imposte effettivamente versate dai contribuenti.

5 Per un approfondimento si veda: Consolini e Donatiello, 2015.

6 Per il confronto con gli aggregati di Contabilità Nazionale, è stata utilizzata una definizione di reddito lordo EU-SILC "allargata" alle componenti che, secondo il Regolamento europeo dell'indagine, non sono ancora incluse nella definizione della variabile di reddito al lordo delle imposte e dei contributi sociali (HY010), ossia sono stati inclusi gli affitti imputati, tutti i fringe benefits, l'autoconsumo e i contributi sociali posti a carico del datore di lavoro.

6.3 La parametrizzazione dei contributi sociali nel modello SM2-EU-SILC

Pur utilizzando congiuntamente i dati campionari e le dichiarazioni fiscali per la produzione dei redditi lordi nell'indagine EU-SILC, i contributi sociali dei lavoratori dipendenti, autonomi, parasubordinati e dei datori di lavoro sono stimati dal modello SM2-EU-SILC (Donatiello, Betti e Consolini, 2012). La parziale copertura delle fonti amministrative attualmente disponibili, in particolare per la contribuzione degli autonomi, e il confronto con fonti esterne di benchmark hanno suggerito, fin dall'anno di indagine 2008, di utilizzare esclusivamente la stima da modello per il calcolo dei contributi sociali.

L'indagine EU-SILC stima i contributi sociali effettivi, ossia i contributi obbligatori (e quelli volontari se previsti dai contratti collettivi di lavoro) per il conseguimento delle prestazioni previdenziali e assistenziali (malattia, invalidità, malattie professionali o infortuni sul lavoro, vecchiaia, maternità), posti a carico dei lavoratori dipendenti, autonomi, parasubordinati (collaboratori coordinati e continuativi e collaboratori a progetto) e dei datori di lavoro. I contributi figurativi non sono attualmente stimati nell'indagine EU-SILC.

In Italia, le aliquote contributive sul reddito da lavoro sono differenziate per fonte di reddito, qualifica professionale e settore di attività economica. I contributi previdenziali dei lavoratori dipendenti sono calcolati sulla retribuzione lorda con importi minimi e massimi di contribuzione, questi inoltre variano in base alla dimensione dell'impresa, in termini di numero di addetti, al settore di attività economica e alla condizione professionale del lavoratore (Tavola 6.2). Nel modello SM2-EU-SILC sono applicate le aliquote contributive previste per i diversi settori di attività economica distinte in base a tre categorie professionali, ossia operai, impiegati e dirigenti. Tuttavia, le informazioni rilevate non consentono di distinguere tra lavoratori dipendenti dell'industria e delle imprese artigiane e per questi ultimi sono applicate le aliquote previste per l'industria. I contributi sociali posti a carico del datore di lavoro sono stimati in SM2-EU-SILC come somma dei contributi effettivi del datore di lavoro e dell'onere per l'accantonamento al trattamento di fine rapporto[7] (Tfr).

Secondo il regolamento EU-SILC, i contributi sociali dovuti sul reddito da lavoro autonomo includono sia i contributi sociali dei lavoratori autonomi sia i contributi dei lavoratori parasubordinati, a cui sono aggiunti i contributi posti a carico dei committenti. Le aliquote contributive previste dall'ordinamento per i lavoratori autonomi sono distinte per gli artigiani e i commercianti, i lavoratori autonomi agricoli e i liberi professionisti.

Il modello di micro-simulazione calcola le aliquote contributive per artigiani e commercianti differenziate per categoria, fasce di reddito ed età del lavoratore. La base imponibile per gli artigiani e commercianti presenta un minimale e massimale aggiornato annualmente (ad esempio nel 2015, euro 15.548 e 76.872) e va considerato che in caso di reddito da lavoro autonomo inferiore al minimale il lavoratore paga comunque la contribuzione prevista per il minimale di reddito (Tavola 6.3).

Nel settore agricolo, l'importo del contributo annuo dovuto dai coltivatori diretti, dai coloni e dai mezzadri e dagli imprenditori agricoli professionali è calcolato sulla base del reddito agrario corrispondente ad una delle quattro fasce di reddito stabilite dalla legge e alle giornate lavorative convenzionali necessarie alla conduzione del fondo (Tavola 6.4). La base imponibile per ogni fascia di reddito è determinata moltiplicando il reddito

7 La quota di accantonamento del TFR è stimata per tutti i lavoratori dipendenti. Le informazioni disponibili nell'indagine non consentono di distinguere i dipendenti del settore privato che hanno deciso di destinare il TFR ai Fondi Pensione.

medio convenzionale (stabilito annualmente con decreto del Ministero del lavoro e della previdenza sociale e che per l'anno 2015 è pari a 55,05 euro) per il numero delle giornate lavorative convenzionali. Nel modello SM2-EU-SILC, al reddito agrario si applicano le aliquote contributive previste e ridotte per i soggetti assicurati di età inferiore ai 21 anni (Tavola 6.5).

La categoria dei liberi professionisti include i soci di impresa e i lavoratori autonomi (imprenditore, titolare e coadiuvante di un'impresa familiare) suddivisi in: (a) liberi professionisti non iscritti a nessun altro fondo di assicurazione sociale obbligatoria e quindi iscritti alla Gestione Separata; (b) liberi professionisti iscritti a qualsiasi altro fondo di assicurazione sociale obbligatoria che versano contributi supplementari, nonché i lavoratori autonomi occasionali, se il loro reddito annuo lordo da lavoro autonomo supera i 5.000 euro. La categoria (a) comprende anche i dottori di ricerca o i beneficiari di borse di studio, oltre ai lavoratori parasubordinati, che presentano una condizione lavorativa che è essenzialmente intermedia tra occupazione dipendente e indipendente. I parasubordinati sono considerati lavoratori autonomi, ma hanno un trattamento fiscale assimilato al reddito da lavoro dipendente e, per questo motivo, i contributi sociali sono pagati anche dal committente. Negli ultimi anni, i contributi sociali a carico dei parasubordinati sono aumentati progressivamente fino ad avvicinarsi alle aliquote dei lavoratori dipendenti. Nel 2015, per i professionisti iscritti alla Gestione Separata, l'aliquota contributiva è pari al 30,72% con un massimale di reddito annuo di 100.324 euro. Poiché le informazioni dell'indagine non consentono di distinguere gli iscritti alla Gestione Separata, questa aliquota, nel modello, viene applicata soltanto a coloro che dichiarano di avere unicamente redditi derivanti da lavoro parasubordinato e non presentano alcun altro tipo di reddito o pensione. Per gli altri liberi professionisti viene applicata l'aliquota del 23,5 per cento del reddito imponibile con il massimale annuo di 100.324 euro (Tavola 6.6).

Le stime finali del modello SM2-EU-SILC dei contributi sociali obbligatori a carico dei lavoratori e dei datori di lavoro sono poi sottoposte ad una procedura di validazione utilizzando le informazioni longitudinali e le fonti esterne disponibili, quali la Contabilità Nazionale e l'Inps[8].

8 Si veda Inps, 2015.

Tavola 6.2 - Aliquote contributive di lavoratori dipendenti del settore privato incluse nel modello SM2/EU-SILC per settore di attività economica, qualifica professionale e numero di addetti. Anno 2015 (valori percentuali)

SETTORE DI ATTIVITÀ ECONOMICA	QUALIFICA PROFESSIONALE	FINO A 15		DA 16 A 50		PIÙ DI 50	
		DATORE DI LAVORO	LAVORATORE	DATORE DI LAVORO	LAVORATORE	DATORE DI LAVORO	LAVORATORE
Industria in genere	Operai	30,88	9,19	31,78	9,49	32,08	9,49
	Impiegati	28,66	9,19	29,56	9,49	29,86	9,49
	Dirigenti	26,96	9,19	30,06	9,19	27,26	9,19
Costruzioni	Operai	34,98	9,19	35,58	9,49	35,58	9,49
	Impiegati	29,46	9,19	30,06	9,49	30,36	9,49
	Dirigenti	26,96	9,19	26,96	9,19	26,96	9,19
Industria estrattiva	Operai	32,68	9,19	33,58	9,49	33,58	9,49
	Impiegati	28,66	9,19	29,56	9,49	29,86	9,49
	Dirigenti	26,96	9,19	27,26	9,19	27,26	9,19
Attività finanziarie e assicurative (a)	Operai	26,76	9,19	26,76	9,19	26,76	9,19
	Impiegati	26,76	9,19	26,76	9,19	26,76	9,19
	Dirigenti	26,76	9,19	26,76	9,19	26,76	9,19
Commercio all'ingrosso e al dettaglio	Operai	28,98	9,14	28,98	9,14	29,88	9,49
	Impiegati	28,98	9,14	28,98	9,14	29,88	9,49
	Dirigenti	26,54	9,19	26,54	9,19	26,84	9,19
Servizi in genere (a)	Operai	29,48	9,19	29,48	9,19	29,48	9,19
	Impiegati	29,48	9,19	29,48	9,19	29,48	9,19
	Dirigenti	26,54	9,19	26,54	9,19	26,54	9,19
Agricoltura, silvicoltura e pesca (a)	Impiegati	25,63	8,84	25,63	8,84	25,63	8,84
	Dirigenti	24,13	8,84	24,13	8,84	24,13	8,84

Fonte: Inps
(a) Aliquote non differenziate per numero di addetti.

Tavola 6.3 - Aliquote contributive di artigiani e commercianti per classe di reddito e età. Anno 2015 (valori in euro e percentuali)

CLASSE DI REDDITO		ARTIGIANI		COMMERCIANTI	
DA	FINO A	FINO A 21ANNI	PIÙ DI 21 ANNI	FINO A 21ANNI	PIÙ DI 21 ANNI
15.548	46.123	19,74	22,74	19,65	22,65
46.123	76.872	20,74	23,74	20,65	23,65

Fonte: Inps

Tavola 6.4 - Fasce di Reddito dei lavoratori agricoli, giornate lavorative e reddito medio convenzionali. Anno 2015

FASCIA	GIORNATE	REDDITO CONVENZIONALE GIORNALIERO (IN EURO)
1	156	55,05
2	208	
3	256	
4	312	

Fonte: Inps

Tavola 6.5 - Aliquote contributive di artigiani e commercianti per classe di reddito e età. Anno 2015 (valori in euro)

CLASSE DI REDDITO			ETÀ	
FASCIA	DA	FINO A	FINO A 21ANNI	PIÙ DI 21 ANNI
1	0	8.587,80	2.751,09	2.836,97
2	8.587,80	11.450,40	3.375,14	3.489,64
3	11.450,40	14.313,00	3.999,18	4.142,31
4	14.313,00	17.175,60	4.623,23	4.794,99

Fonte: Inps
(a) Coltivatori diretti, mezzadri, coloni e imprenditori agricoli.

Tavola 6.6 - Aliquote contributive per liberi professionisti e lavoratori parasubordinati (a). Anno 2015 (valori percentuali)

	ALIQUOTE
Iscritti alla Gestione separata	30,72
Iscritti ad altre gestioni	23,5

Fonte: Inps
(a) Massimale di reddito annuo: euro 100.324.

6.4 Le imposte micro-simulate

Il processo produttivo dell'Indagine EU-SILC ha attualmente a disposizione le informazioni amministrative derivanti dalle dichiarazioni fiscali (Modello 730, Unico Persone Fisiche, Certificazione Unica) e dal Casellario pensionistico, le quali assieme alle informazioni rilevate dall'indagine[9] concorrono alla stima delle variabili di reddito EU-SILC.

La Tavola 6.7 presenta la copertura delle informazioni derivante da archivio fiscale per gli individui (26.691) a cui è stata stimata l'imposta nell'indagine EU-SILC 2015. Il 94,4 per cento di questi individui è presente in almeno uno degli archivi disponibili, mentre per il restante 5,6 per cento non si ha alcuna informazione.

Per gli individui che presentano un'imposta, le "tasse amministrative", ossia derivanti dalle dichiarazioni fiscali o, nel caso delle pensioni, dal Casellario pensionistico, costituiscono l'83,2 per cento del totale, includendo l'1,1 per cento di lavoratori autonomi con un'imposta stimata pari a zero[10] (Tavola 6.8). Il restante 16,8 per cento delle imposte deriva dal modello di micro-simulazione che stima, quindi, le imposte per gli individui non agganciati negli archivi, poiché si ipotizza una sotto copertura degli archivi stessi, e per gli individui presenti negli archivi fiscali ma le cui informazioni sulle imposte non vengono utilizzate. Si tratta, ad esempio, di imposte per tipologie di reddito non incluse nella definizione delle variabili target EU-SILC[11], oppure di imposte non coerenti o affette da errori o non conformi alle procedure di controllo e correzione utilizzate nel processo produttivo dei redditi lordi dell'indagine.

Le imposte che derivano dagli archivi fiscali sono, infatti, sottoposte ad una complessa procedura di trattamento dei dati che prevede la riconciliazione delle informazioni riguardanti lo stesso individuo tra i diversi archivi e la successiva integrazione di queste informazioni con i dati raccolti dall'indagine. Dopo la fase di integrazione, e ai fini dell'utilizzo delle imposte fiscali per la costruzione dei redditi lordi, si procede ad un controllo di coerenza e correzione dei dati netti, lordi, delle ritenute e imposte di fonte amministrativa, per elimi-

Tavola 6.7 - Copertura degli archivi amministrativi per i redditi lordi dell'indagine EU-SILC 2015 (redditi 2014)

	EU-SILC	VALORE PERCENTUALE
Campione effettivo	42.987	
Intervistati (a)	36.602	
Individui con imposta	26.691	100,0
Agganciati negli archivi	25.204	94,4
Non agganciati negli archivi	1.487	5,6

(a) Individui con 16 anni e più

9 Sulla metodologia di integrazione si veda Consolini, 2009.

10 Secondo il Regolamento dell'indagine, le imposte dei lavoratori dipendenti e dei pensionati corrispondono alla ritenuta alla fonte. Per i lavoratori autonomi, diversamente da quanto previsto dal Regolamento, la tassazione corrisponde all'imposta netta poiché in Italia gli acconti e le ritenute sul reddito da lavoro autonomo, a differenza dell'imposta anticipata dei dipendenti e pensionati, possono divergere anche significativamente dall'imposta finale pagata.

11 Come ad esempio per i guadagni derivanti da *capital gains*.

nare eventuali anomalie tra ritenute, imposte finali e redditi corrispondenti. In questa fase, alcuni dati amministrativi risultano inidonei alla stima dei redditi lordi dell'indagine e quindi vengono utilizzate le stime derivanti dal modello di micro-simulazione.

Tavola 6.8 - Imposte per tipologia nell'indagine EU-SILC 2015 (redditi 2014)

	EU-SILC	VALORE PERCENTUALE
Individui con imposta	26.691	100
Agganciati negli archivi e con imposte da registro	21.910	82,1
Agganciati negli archivi e con imposte da registro pari a 0	294	01:01
Agganciati e non negli archivi e con imposte microsimulate	4.487	16,8

Per capire come l'utilizzo delle imposte amministrative, per oltre l'83 per cento del totale, modifichi la distribuzione delle imposte per tipologia di reddito è stato fatto un confronto tra le stime finali delle imposte dell'indagine EU-SILC 2015 (le quali sono il risultato dell'utilizzo congiunto delle tasse amministrative e delle tasse micro-simulate) e le stime da modello (ossia senza l'integrazione con le imposte fiscali[12]).

Occorre tenere presente che il modello SM2-EU-SILC stima le imposte e i contributi sociali operando una conversione dei redditi netti in lordi. In presenza di redditi netti amministrativi superiori ai redditi netti rilevati, i redditi netti e lordi dei micro-dati finali derivano dagli archivi fiscali e conseguentemente le stime finali non differiscono molto dai dati amministrativi[13]. Quando invece i redditi netti rilevati risultano superiori ai redditi netti fiscali, i redditi netti dei micro-dati finali sono quelli rilevati mentre le imposte, come riportato in precedenza, derivano dagli archivi e quindi i redditi netti e lordi finali differiscono da quelli amministrativi[14].

L'applicazione di questo metodo ha come conseguenza di non utilizzare fattori di correzione delle imposte per tenere conto dell'evasione fiscale[15], poiché oltre l'83 per cento delle imposte deriva dal fisco e non necessita di alcun aggiustamento. Inoltre, le procedure di validazione con le fonti esterne di benchmark consentono di valutare appropriatamente i totali delle imposte, inclusa la quota micro-simulata.

Al fine di valutare l'impatto delle imposte amministrative sulla stima dei redditi lordi 2014 sono state analizzate e poste a confronto le stime micro-simulate e quelle finali. In questo studio, per procedere ad un confronto corretto, è stato necessario operare alcuni aggiustamenti sui dati micro-simulati. Per prima cosa sono state corrette le stime micro-simulate per tenere conto dell'evasione fiscale ed è stata concentrata l'attenzione su alcune categorie professionali in cui, notoriamente, è più frequente la mancanza di dichiarazioni fiscali come, ad esempio, i lavoratori del settore delle costruzioni o dell'agricoltura, i lavoratori domestici, i camerieri, gli operatori sanitari, eccetera. Per operare il confronto su basi omogenee, alle imposte microsimulate sono stati poi aggiunti i saldi fiscali (credito o debito) dichiarati all'intervista. Dopo le correzioni, sono

12 Lasciando invariati i dati di input del modello.

13 Vi è una differenza dovuta all'inserimento di una componente stocastica sulle imposte, per motivi di anonimizzazione dei dati utilizzati.

14 Vi sono molti studi in letteratura sulla stima dell'evasione fiscale basata sulla differenza tra redditi dichiarati ad un'indagine campionaria e redditi derivanti dalle dichiarazioni fiscali. Tuttavia, considerando che la stima dell'evasione fiscale esula dagli obiettivi di produzione dell'indagine EU-SILC e che le informazioni disponibili non consentono di distinguere tra elusione fiscale, realizzata in conformità a norme previste dall'ordinamento, ed evasione, fino ad ora non abbiamo effettuato studi in questa direzione.

15 Tali fattori di correzione sono solitamente utilizzati in tutti i modelli di micro-simulazione.

state ricalcolate le imposte micro-simulate per tipologia di reddito e comparate con le imposte finali (Grafico 6.1).

Come atteso, l'utilizzo delle imposte derivanti dalle dichiarazioni fiscali ha un impatto più evidente sulla distribuzione dei redditi lordi da lavoro autonomo, dove le stime micro-simulate sono più elevate per tutti gli scaglioni di reddito. È molto probabile che gli autonomi, in sede di dichiarazione dei redditi, riescano ad ottenere maggiori abbattimenti degli imponibili rispetto a quanto simulato dal modello in termini di deduzioni e detrazioni d'imposta. Anche nel caso delle imposte sui redditi da pensione, le stime micro-simulate sono superiori per quasi tutti gli scaglioni di reddito e in particolare per le due classi inferiori. Come risulta evidente dal grafico, questo andamento è determinato dalle pensioni da lavoro, poichè le *imposte finali* (amministrative) sulle pensioni da lavoro includono una tassazione più bassa per le pensioni di invalidità pagate dopo i 65 anni le quali, secondo il Regolamento dell'indagine, confluiscono in questa tipologia di pensione. Il modello, invece, stima un'aliquota media per tutte le pensioni da lavoro, incluse quelle di invalidità, che confluiscono nella stessa variabile target. Nel caso dei redditi da lavoro dipendente, le stime micro-simulate sono simili alle stime finali solo per gli scaglioni di reddito più elevati (dai 28.000 euro in poi). Infine, le differenze rilevate tra stime da modello e incidenze finali risultano meno marcate confrontando il totale delle imposte.

Figura 6.1 - Incidenza delle imposte finali e simulate per fonte e scaglioni di reddito. Indagine EU-SILC 2015 (redditi 2014)

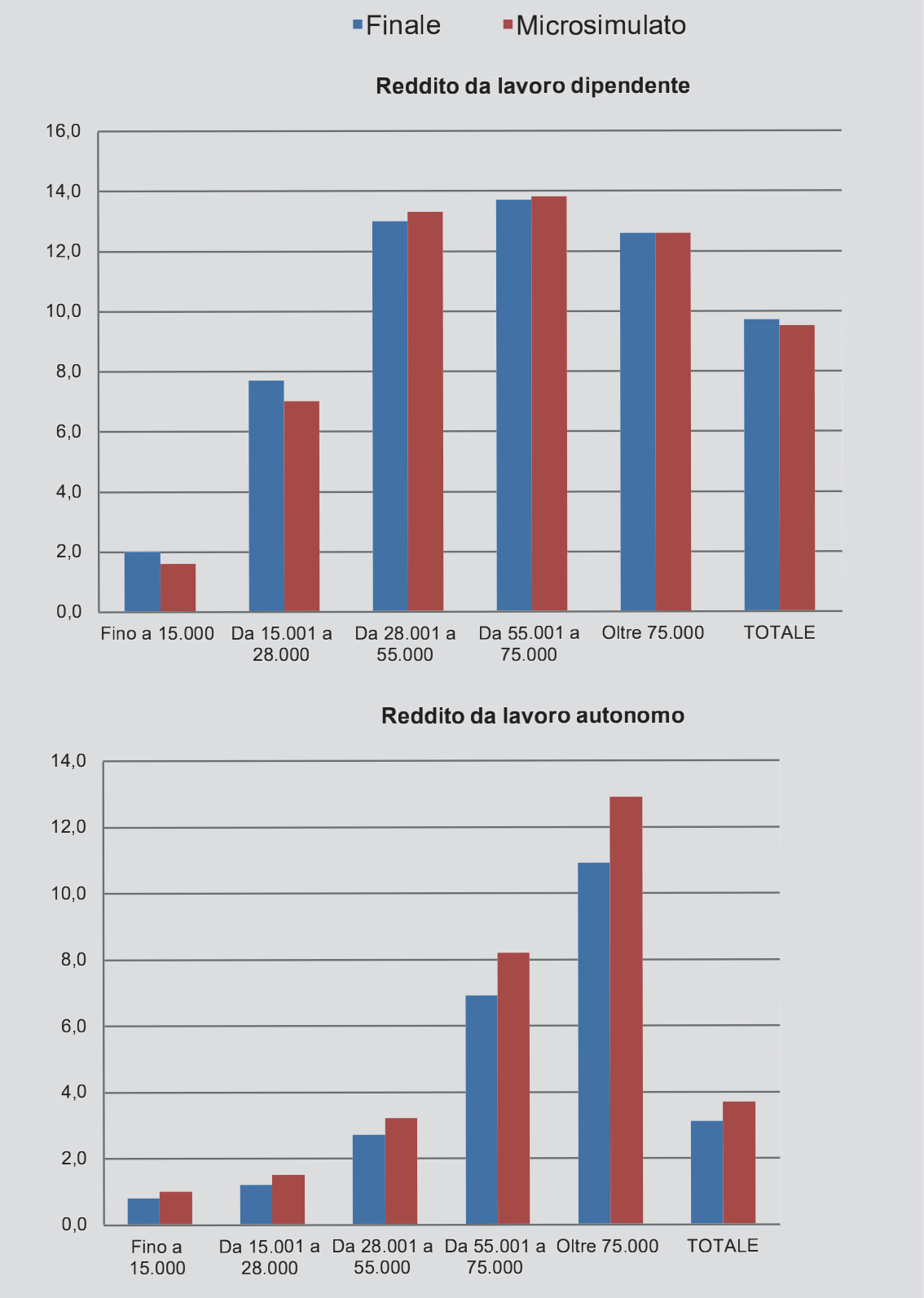
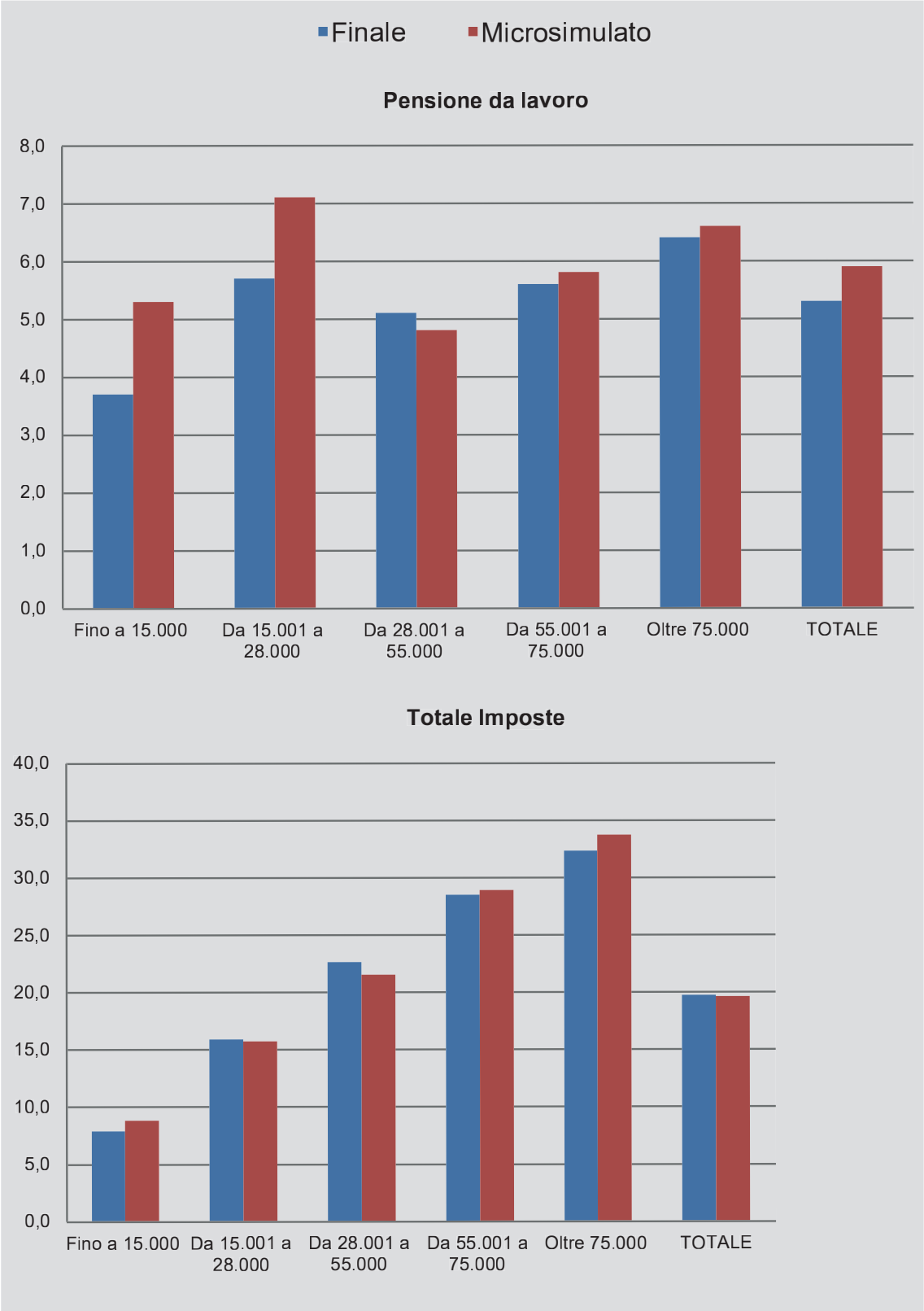


Figura 6.1 segue - **Incidenza delle imposte finali e simulate per fonte e scaglioni di reddito. Indagine EU-SILC 2015** (redditi 2014)



6.5 Conclusioni

La metodologia consolidata di produzione multi-fonte dei redditi dell'indagine EU-SILC, che utilizza congiuntamente le informazioni amministrative, i dati rilevati e le stime micro-simulate, ha consentito negli anni di accumulare esperienza nel trattamento degli archivi fiscali a fini statistici, di ampliare il numero degli archivi utilizzati e di migliorare la metodologia di integrazione delle diverse fonti con l'intento di fornire dati di qualità e il più possibile esaustivi.

Il processo produttivo dei redditi al lordo dell'imposizione fiscale e contributiva si avvale delle informazioni disponibili negli archivi fiscali, utilizzate e integrate con le stime del modello di micro-simulazione SM2-EU-SILC (Betti, Donatiello e Verma, 2011). Quest'ultimo utilizza parte delle informazioni fiscali nel file di input (ad esempio in termini di oneri deducibili e detraibili) e soprattutto si sfruttano le informazioni fiscali come benchmark per la validazione delle stime micro-simulate. In particolare il confronto tra stime da modello e dati di archivio ha consentito, negli anni, di migliorare le stime micro-simulate e di avvalersi di una comparazione incrociata dei risultati utile anche a individuare incoerenze o errori nei dati amministrativi.

Certamente la disponibilità di dati derivanti dalle dichiarazioni fiscali dei contribuenti e dal Casellario Pensionistico i quali, opportunamente trattati, forniscono oltre i quattro quinti delle imposte finali EU-SILC, rappresenta un indubbio vantaggio per l'accuratezza delle stime. Come risulta dal confronto operato tra stime micro-simulate e stime finali nell'indagine EU-SILC 2015, l'utilizzo delle imposte amministrative modifica di fatto la distribuzione delle imposte finali per tipologia di reddito rispetto alla distribuzione micro-simulata, a conferma che la scelta di utilizzare congiuntamente dati fiscali e micro-simulati consente di migliorare la qualità delle statistiche prodotte.

7. IL PROCESSO DI VALIDAZIONE CON FONTI ESTERNE A SUPPORTO DEL TRATTAMENTO DEI DATI[1]

7.1 Introduzione

La comparazione a livello macro delle stime d'indagine con i dati di fonte esterna è una passo indispensabile per poter certificare la qualità dei dati da diffondere, ma è del pari importante per segnalare eventuali criticità nel processo di trattamento e di stima (identificazione e correzione di dati anomali, valutazione dell'impatto delle ipotesi sull'evasione del reddito e del peso relativo delle componenti microsimulate vs quelle amministrative, calibrazione dei pesi, eccetera). Infatti, laddove si ravvisino significativi e strutturali scostamenti rispetto ai *benchmark*, è inevitabile dare avvio a studi e approfondimenti per individuarne la causa, ciò anche al fine di pianificare interventi di miglioramento nel processo di trattamento dati.

In ogni caso nella valutazione e comparazione con dati da fonte esterna, non si può non tener conto della natura campionaria dell'indagine e quindi della presenza di un errore campionario nelle stime prodotte.

Non mancano in letteratura esempi di utilizzo di dati amministrativi per la correzione di stime campionarie sui redditi, si può citare il lavoro di Briker *et al.*, 2015, che affronta il problema della sotto-rappresentatività delle famiglie che si collocano nella coda superiore dei redditi e della ricchezza.

Nel processo di validazione, la prima fase consiste nell'individuazione delle fonti esterne (statistiche o amministrative) che rilevano la stessa tipologia di informazione. Le successive riguardano l'analisi della copertura delle fonti, l'armonizzazione delle definizioni e più in generale l'analisi dei metadati sui cui si fondano i dati.

Il processo di validazione delle componenti di reddito in EU-SILC si espleta, quindi, confrontando le stime campionarie con le principali fonti esterne che rappresentano il nostro modello di riferimento.

La fase preliminare per il confronto consiste nel determinare l'aggregato massimo omogeneo che permette la comparazione. Le componenti di reddito EU-SILC che sono paragonabili su base omogenea con il dato di fonte esterna comprendono: i redditi pensionistici, il reddito da lavoro dipendente e alcuni trasferimenti non pensionistici. Viceversa nei casi dei redditi autonomi non è possibile identificare un aggregato omogeneo, tra le fonti esterne disponibili, cui confrontare la stima EU-SILC.

1 L'autore di questo capitolo è Paolo Consolini.

7.2 La validazione dei redditi pensionistici

Il reddito pensionistico è misurato come sommatoria dei singoli trattamenti che lo compongono in senso stretto, cioè al netto di altre componenti, quali ad esempio il TFR erogato al momento della quiescenza.

La fonte principale per la rilevazione a livello macro dei trasferimenti pensionistici in Italia è rappresentata dall'indagine sui bilanci consuntivi degli enti previdenziali che registra i flussi monetari generati dal centro erogatore di spesa[2].

Il dato di cassa riportato nei bilanci (lato centro erogatore) è quello che più si accosta al concetto di reddito da pensione trasferito alle famiglie/individui di EU-SILC (lato beneficiari), ovvero ai pagamenti ricevuti nell'anno solare. Le voci che compongono la spesa pensionistica dei bilanci di cui sopra comprendono: le pensioni di invalidità, di vecchiaia e ai superstiti (previdenziali), le rendite per infortuni (previdenziali), le pensioni o assegni sociali (assistenziali), le pensioni e indennità a invalidi civili, non vedenti e non udenti (assistenziali) e i prepensionamenti (in parte di natura previdenziale e in parte a carico della fiscalità generale). Gli enti previdenziali sono tanto quelli afferenti al regime di base, erogatori di prestazioni sociali in base alle norme dell'assicurazione generale obbligatoria, quanto quelli appartenenti al regime complementare che forniscono prestazioni aggiuntive o integrative rispetto al regime di base. L'indagine sui bilanci riporta per l'anno 2015 un dato di spesa per i trasferimenti lordi[3] pensionistici pari a 280,7 miliardi di euro in termini di cassa, a fronte di 281,2 miliardi di euro in impegni (Tavola 1).

Tavola 1.1 – Le spese pensionistiche da indagine sui bilanci consuntivi degli enti previdenziali. Anno 2015

VOCI	Impegni (migliaia euro)	Pagamenti (migliaia euro)
SPESE PER PRESTAZIONI SOCIALI	**317.885.223**	**317.488.123**
di cui pensionistiche:	***281.156.624***	***280.741.111***
- Pensioni di invalidità, vecchiaia e superstiti	252.861.032	252.647.098
- Assegni sociali ad ultra 65 enni	4.727.515	4.723.250
- Pensioni e indennità a invalidi civili, non vedenti e non udenti	16.885.563	16.688.788
- Rendite per infortuni o assegni vitalizi	4.569.559	4.569.489
- Prepensionamenti(L.223/91)	1.477.811	1.477.811

La spesa (lorda) pensionistica trasferita alle famiglie da fonte EU-SILC è stimata in 273,5 miliardi di euro. Il reddito lordo pensionistico assegnato a ciascun pensionato nell'indagine è frutto dell'abbinamento e dell'integrazione di più fonti amministrative, quali il Casellario pensionistico, gli archivi fiscali (CU, 730 e Upf) e le informazioni d'indagine (Consolini, 2009). Si fa presente che il ricorso a tecniche di rilevazione multi-fonte per le pensioni trova giustificazione nella presenza di lacune informative che impediscono la piena copertura tramite l'utilizzo di una fonte esclusiva. Nella riprogettazione dell'edizione 2016 (redditi 2015), almeno sul fronte delle pensioni di base, si è ritenuto di poter considerare sostanzialmente colmato tale deficit[4], imponendo il ricorso alla sola fonte amministrativa e

2 L'indagine è utile ai fini della costruzione del Conto economico delle Amministrazioni pubbliche, elaborato secondo gli schemi contabili del Sistema europeo dei conti economici integrati (Sec2010) nonché del conto economico della protezione sociale, costruito secondo i criteri previsti dal Sistema europeo delle statistiche integrate della protezione sociale (SESPROS).

3 Al lordo delle ritenute fiscali alla fonte.

4 Miglioramenti in termini di qualità dei registri anagrafici e di tempistica della rilevazione sul campo, tale da rendere possibile il recupero dei codici fiscali sui componenti aggiuntivi non presenti nelle liste iniziali (anagrafiche per la prima wave o esito della precedente rilevazione), hanno consentito di raggiungere una percentuale di successo

riducendo, in questo modo, il carico sui rispondenti. Si è invece mantenuta la tecnica multifonte (amministrativa e d'indagine) per le pensioni appartenenti al regime complementare (fondi negoziali) in quanto parzialmente coperte dai registri e archivi esistenti.

Sebbene la definizione di "trasferimento lordo pensionistico" risulti omogenea tra la fonte EU-SILC e l'indagine dei bilanci consuntivi, in realtà i due aggregati non sono perfettamente comparabili in termini di popolazione di riferimento (copertura). Nei bilanci figurano i trattamenti a favore dei beneficiari minori di 16 anni, dei pensionati residenti in convivenze (case di riposo, case di cura, ospedali, ecc.) e dei residenti all'estero che esulano dal campo di osservazione di EU-SILC. Quest'ultima, dal canto suo, include (a partire dall'anno 2015) le pensioni erogate dagli enti previdenziali esteri ai residenti in Italia che, per definizione, non sono registrati nell'indagine sui bilanci degli enti nazionali di previdenza. Non potendo identificare, all'interno dei bilanci previdenziali, i diversi collettivi di pensionati secondo l'età o la residenza, il confronto su base omogenea tra dato amministrativo e rilevazione statistica non è pienamente realizzabile.

Il Casellario pensionistico costituisce una fonte di riferimento alternativa per la stima macro della spesa pensionistica e, sotto certi aspetti, consente di migliorare la comparabilità con le stime di EU-SILC. La fonte in questione rileva l'importo lordo mensile erogato dall'ente previdenziale al 31 dicembre dell'anno di riferimento dei redditi, comprensivo delle seguenti componenti: importo base, incremento collegato alla variazione dell'indice del costo della vita e alla dinamica delle retribuzioni, tredicesima mensilità ed eventuali altri assegni e arretrati. La spesa pensionistica da Casellario è definita come dato di stock e pertanto non coincide col dato di flusso (annuo) riportato nei valori contabili dei bilanci degli enti previdenziali. Un'approssimazione del flusso annuo dei trasferimenti lordi pensionistici si ottiene moltiplicando, per ciascun trattamento e per il totale dei beneficiari, l'importo relativo alla mensilità di dicembre nell'anno di riferimento dei redditi (nell'esempio il 2015) per i 12 mesi del calendario, con l'inclusione dell'eventuale tredicesima. L'importo annuo della spesa pensionistica calcolata da Casellario si approssima all'ammontare di spesa effettivamente erogata dagli enti previdenziali nella misura in cui l'errore dovuto alla mancata rilevazione dei trattamenti cessati nell'anno corrente (non registrati a fine anno) si compensa con l'errore associato alla sovrastima degli importi annui sui nuovi trattamenti (per convenzione ad essi si assegna un valore di 12 mensilità a fronte di un numero effettivo che dipende in realtà dal mese di decorrenza). Considerando l'universo dei percettori di pensione da Casellario al 31 dicembre 2015 (16,1 milioni), si desume una spesa lorda pensionistica pari a 282,4 miliardi di euro[5]. Se, ai fini della comparabilità con EU-SILC, si restringe il campo di osservazione ai soli residenti in Italia e ai percettori di 16 anni e più, si determina un valore di spesa pensionistica da Casellario pari a 279,7 miliardi di euro per un totale di 15,3 milioni di pensionati.

La terza fonte, infine, è rappresentata dal Conto economico consolidato della protezione sociale di contabilità nazionale, la quale utilizza come fonti l'indagine sui bilanci degli enti previdenziali e altri dati finanziari per settore istituzionale. All'interno del conto satellite l'importo complessivo delle prestazioni di tipo pensionistico è pari a 283,3 miliardi di euro, un valore di poco superiore rispetto al dato da Casellario e alla spesa desumibile dai bilanci degli enti previdenziali.

superiore al 99% nell'abbinamento delle unità d'indagine con gli archivi amministrativi.

5 L'ammontare di spesa supera di circa 1,7 miliardi di euro il dato relativo ai pagamenti pensionistici dell'indagine sui bilanci consuntivi. Lo scarto è dovuto in parte alle differenze tra le definizioni di "reddito lordo pensionistico annuo" e in parte al fatto che il collettivo degli enti previdenziali rilevati nel Casellario è più ampio rispetto all'indagine sui bilanci.

7.3 La validazione dei redditi da lavoro dipendente

Il reddito lordo da lavoro dipendente rappresenta il secondo grande aggregato di reddito soggetto a validazione. La stima campionaria EU-SILC viene pertanto confrontata a livello "macro" con le fonti *benchmark* disponibili (amministrative e statistiche).

Il dato statistico ufficiale sull'ammontare delle "retribuzioni lorde" è fornito nell'ambito della Contabilità Nazionale dal conto sul reddito disponibile delle famiglie italiane (risorse a disposizione del settore Istituzionale Famiglie). In base a questa fonte l'ammontare complessivo delle retribuzioni lorde è pari a 482,1 miliardi di euro. Nell'indagine EU-SILC il valore complessivo delle retribuzioni lorde è stimato in 492,8 miliardi di euro. Occorre tuttavia precisare che in questa stima sono incluse una serie di voci economiche non facenti parte della retribuzione che tuttavia, a causa di carenze informative, non sono scorporabili. Tra esse figurano le prestazioni sociali erogate direttamente dai datori di lavoro (ad esempio i primi tre giorni di assenza –carenza– per la malattia dei lavoratori del settore privato, le integrazioni alle retribuzioni nei periodi di maternità prevista da CCNL e così via) e talune anticipate dal datore di lavoro per conto dell'Inps (malattia, maternità, congedi parentali, ecc.). Nei conti economici di contabilità nazionale queste non figurano nelle retribuzioni, ma sono collocate tra le prestazioni sociali. L'ammontare delle prestazioni in questione è pari a circa 6,6 miliardi di euro, che rappresenta la quota parte delle prestazioni sociali erogate direttamente dai datori di lavoro privati senza finalità previdenziali. Se volessimo, quindi, confrontare il dato EU-SILC relativo alle retribuzioni complessive su base omogenea, dovremmo aggiungere 6,6 miliardi di euro ai 482,1 miliardi rilevati dal conto del reddito disponibile, ottenendo così uno scarto dello +0,8% dell'indagine sui redditi rispetto al suo *benchmark*.

A partire dall'edizione 2016 si è sviluppato un nuovo metodo per riallocare la componente cassa integrazione anticipata dal datore di lavoro (cfr. Capitolo 10), separandola dalla retribuzione, con l'uso di soli dati di fonte amministrativa Inps. Ciò ha permesso di affinare la misurazione della stessa componente retributiva. Nelle quattro precedenti edizioni dell'indagine la medesima procedura di riallocazione veniva effettuata sulla base delle sole informazioni riportate durante l'intervista che tuttavia risultavano carenti.

Il confronto tra stime EU-SILC e fonti esterne non si limita all'aggregato della retribuzione lorda ma prevede un'analisi comparativa delle varie sotto-componenti. In particolare si è proceduto ad effettuare una ricognizione delle fonti esterne che rilevano le imposte sul reddito da lavoro dipendente e i contributi sociali a carico dei lavoratori subordinati.

Il dato sulle imposte sui redditi da lavoro dipendente è rilevabile dalla fonte MEF che pubblica le principali poste delle dichiarazioni dei redditi, distinte anche in base alla tipologia di contribuente. In particolare il valore dell'imposta si ottiene sommando il valore dell'Irpef (imposta netta) e le addizionali regionali e comunali applicate alla quota dei redditi da lavoro dipendente. Nel 2015 esse ammontano rispettivamente a 92,9 e 9,8 miliardi di euro sul reddito complessivo del collettivo dei percettori di reddito da lavoro dipendente da cui, tolte le componenti non riferibili alla retribuzione, si ricava un valore d'imposta pari 93,6 miliardi di euro. In EU-SILC si stima per lo stesso anno un totale di 92,6 miliardi di imposte dirette sui redditi da lavoro dipendente, con un minimo scarto rispetto al *benchmark* (0,7%).

I contributi sociali per la quota a carico dei lavoratori dipendenti, rilevati nel conto satellite della protezione sociale di contabilità nazionale, ammontano a 42,8 miliardi di euro contro i 43 miliardi stimati da EU-SILC. In definitiva, nel caso della retribuzione si raggiunge un elevato grado di accostamento tra le stime d'indagine e il dato macro di fonte esterna anche in relazione alle varie sotto-voci che la compongono.

7.4 La validazione dei trasferimenti non pensionistici

La fase di validazione relativa alle prestazioni non pensionistiche ha riguardato gli assegni al nucleo familiare, il trattamento o liquidazione di fine rapporto, i trattamenti di disoccupazione e la cassa integrazione.

In merito agli assegni familiari, l'indagine EU-SILC prevede l'utilizzo congiunto di informazioni da intervista diretta e da archivi amministrativi. Il motivo di tale strategia si spiega con la presenza di errori di sotto-copertura che limitano l'uso esclusivo di quest'ultima fonte. Infatti in essa non figurano i dipendenti pubblici e i disoccupati beneficiari di tale assegno. La strategia di rilevazione multi-fonte è stata rivista a partire dall'edizione 2016 per alleggerire il carico di risposta dei soggetti già compresi nella fonte amministrativa, filtrando l'apposita sezione del questionario ai soli potenziali beneficiari (dipendenti pubblici e disoccupati) non coperti dalla fonte amministrativa. In base all'indagine EU-SILC 2016 si stima un ammontare di spesa trasferita alle famiglie nel 2015 per assegni al nucleo familiare pari a 5,9 miliardi di euro, valore paragonabile al flusso di trasferimenti associati a questa prestazione riportato nel conto satellite della protezione sociale, pari a 6,2 miliardi di euro.

La rilevazione delle liquidazioni di fine rapporto in EU-SILC segue lo stesso ragionamento applicato alle pensioni ovvero, a decorrere dall'edizione 2016, le relative informazioni provengono da sola fonte amministrativa (Certificazioni Uniche del fisco). L'innovazione di processo si giustifica anche in questa circostanza col miglioramento della copertura della fonte a seguito di strategie più efficaci di recupero delle chiavi di aggancio. In base all'indagine EU-SILC 2016 l'ammontare delle liquidazioni di fine rapporto versate nel 2015 è pari a 17,3 miliardi di euro, importo comunque decisamente inferiore al valore complessivo rilevato nei conti della protezione sociale, 23 miliardi di euro. Il motivo di tale scostamento è strettamente connesso al trattamento statistico delle unità 'influenti'[6] (*winsorizing*), opportuno nelle indagini campionarie per migliorare la correttezza e l'efficienza delle stime ma che ha l'effetto di ridurre sensibilmente il valore del totale e della media della distribuzione.

I trattamenti per la disoccupazione, al pari delle pensioni, si rilevano incrociando i dati di fonte previdenziale (archivi Inps), di fonte fiscale (Certificazioni uniche) e, sino all'edizione EU-SILC 2015, da interviste dirette. Anche per queste voci di reddito, infatti, si è deciso di rinunciare, dal 2016, alla rilevazione diretta, grazie alla maggiore copertura delle fonti amministrative, dovuta alla più efficace tecnica di recupero dei codici di abbinamento e all'utilizzo di nuove fonti esterne per includere componenti del trattamento di disoccupazione non rilevate in precedenza. In particolare, come già detto in relazione alle retribuzioni, si è fatto ricorso alla fonte Uniemens dell'Inps per catturare il dato sulla cassa integrazione erogata in busta paga dal datore di lavoro (cfr. Capitolo 10).

I trattamenti di disoccupazione sono distinti in due sottogruppi a seconda che si tratti di disoccupazione parziale (cassa integrazione) o totale (ASpI, NASpI[7], indennità di mobilità), cioè a seconda che vi sia o meno una rescissione del rapporto di lavoro. La fonte deputata per il confronto con le stime EU-SILC sui trattamenti di disoccupazione è l'indagine

6 Per evitare eventuali effetti distorsivi di unità statistiche "influenti", in particolare nelle stime a livello regionale, i valori di alcune variabili di reddito situati nelle code della distribuzione (fino al 5° percentile e oltre il 95° percentile) e con un peso situato nella coda destra della distribuzione regionale dei pesi (oltre l'80° percentile) vengono opportunamente riposizionati: i valori situati nella coda destra vengono riposizionati in funzione del peso tra il 90° e il 95° percentile mentre i valori situati nella coda sinistra vengono riposizionati casualmente tra il 5° e il 10° percentile.

7 L'ASpI, in vigore dal 1 gennaio 2013 in sostituzione dell'indennità di disoccupazione ordinaria non agricola, è l'acronimo di Assicurazione Sociale per l'Impiego; la NASpI, in vigore dal 1 maggio 2015 in sostituzione di ASpI e Mini-ASPI (ASpI con requisiti ridotti), è l'acronimo di Nuova Assicurazione Sociale per l'Impiego.

sui bilanci consuntivi degli enti previdenziali. Dai dati di bilancio figura che nel 2015 l'Inps ha erogato 2,5 miliardi di euro in integrazioni salariali e 11,7 miliardi in sussidi di disoccupazione, contro un valore stimato sui dati EU-SILC rispettivamente pari a 2 miliardi di euro e 9,3 miliardi per le rispettive prestazioni. Il grado di accostamento tra le stime d'indagine e il dato macro di fonte amministrativa sui trattamenti di disoccupazione è da ritenersi meno soddisfacente rispetto ad altre componenti di reddito. A questa prestazione sarà necessario dedicare in futuro degli approfondimenti metodologici volti a spiegare le ragioni di tale distorsione e adottare gli opportuni accorgimenti per migliorare la qualità del dato.

7.5 Conclusioni

La fase di validazione delle stime a livello macro applicata ai tre grandi aggregati di reddito EU-SILC (redditi da lavoro dipendente, trasferimenti pensionistici e non) dimostra l'elevata qualità dell'informazione prodotta e la bontà dell'approccio seguito che prevede, in taluni casi, la sostituzione dei dati di fonte amministrativa a quelli ottenuti tramite intervista diretta e, in altri, l'uso congiunto delle due fonti. Con l'ampliamento delle basi amministrative a disposizione (si pensi ad esempio alle fonti Uniemens e al Casellario degli attivi dell'Inps, o alla banca dati fiscale sugli immobili del MEF) il processo di validazione potrà incorporare nuovi livelli di analisi ai fini della validazione incrociata di variabili comuni. Infine, l'utilizzo di fonti esterne, oltre che essere di ausilio per la revisione e la validazione delle ipotesi insite nella costruzione dei dati, è da considerarsi particolarmente importante in fase di individuazione di errori sia di copertura che di distorsione per mancata risposta. Sotto questo punto di vista, il loro impiego anche in termini di variabili ausiliarie nella fase di calibrazione dei coefficienti di riporto all'universo e/o di stratificazione di un nuovo disegno campionario, potrà consentire, in futuro, di migliorare ulteriormente la qualità delle stime finali.

PARTE TERZA

8. LE FONTI AMMINISTRATIVE PER L'INTEGRAZIONE DEI DATI D'INDAGINE[1]

8.1 Introduzione

Il progetto EU-SILC di rilevazione e misurazione delle variabili di reddito, sviluppatosi in Istat nel corso degli ultimi decenni, include elementi di originalità che lo contraddistinguono rispetto ai restanti Istituti statistici europei. La peculiarità della metodologia consiste nell'utilizzo combinato di informazioni amministrative e campionarie sui redditi, attraverso una strategia che sfrutta il record linkage tra dati individuali di reddito rilevati, rispettivamente, da indagine campionaria con tecnica Capi/Cati (*computer assisted personal/telephone interview*) e da fonti di natura amministrativa dell'Agenzia delle Entrate e dell'INPS (Consolini, 2009). In campo internazionale non mancano esperienze di integrazione a livello micro tra dati di fonte amministrativa e d'indagine. Nel Regno Unito, ad esempio, esse sono applicate nel settore delle statistiche sulla salute, del mercato del lavoro e transizione al pensionamento (Calderwood and Lessof, 2009). In Italia, come altra sperimentazione e applicazione di tecniche di integrazione di dati di diverse fonti su indagini sociali, si può citare l'indagine pilota sull'analisi dell'evoluzione delle storie lavorative dei giovani della provincia di Trento (Bazzoli *et al.*,2018). Gli esperimenti di micro-integrazione tra fonti amministrative e dati di indagini sociali sono, tuttavia, ancora poco diffusi; ciò a causa delle restrizioni imposte dal legislatore in materia di tutela della privacy per il trattamento di dati personali ai fini statistici o di ricerca scientifica (Trivellato, 2017).

I vantaggi che si conseguono da una strategia di raccolta dati multi-fonte, combinata a metodi di micro-integrazione delle informazioni, sono principalmente quattro: riduzione del carico statistico sui rispondenti (*response burden*), migliore livello di copertura del dato, maggiore accuratezza nella misurazione delle variabili di analisi e maggiore dettaglio nel contenuto informativo delle informazioni prodotte (Coli, Consolini e D'Orazio, 2016).

Le indagini campionarie utilizzate per la rilevazione di fenomeni sociali (lavoro, condizioni economiche, aspetti della vita quotidiana) sono tipicamente soggette a una molteplicità di errori non campionari (e campionari) che possono inficiare la qualità delle stime finali. L'incapacità dell'intervistato di quantificare con precisione eventi distanti nel tempo o semplicemente di ricordarne l'esistenza (in linea teorica potrebbero trascorrere 12 mesi dal momento dell'intervista all'anno di riferimento dell'indagine), l'effetto di interazione intervistato-intervistatore, la reticenza a dichiarare il vero, costituiscono alcuni degli esempi dei potenziali errori insiti nell'indagini campionarie che utilizzano come unico strumento di rilevazione l'intervista. D'altra parte, il ricorso all'uso di risposte *proxy* (quando a rispondere è un soggetto diverso rispetto all'unità di osservazione), seppur limitato in un'indagine come quella di EU-SILC, non può che compromettere la qualità dell'informazione delle unità statistiche coinvolte. A ciò si aggiungono tipologie di errori di natura non campionaria, come quelli imputabili a registrazione, codifica, auto-selezione dei rispondenti, eccetera. In questo contesto, la possibilità di integrare basi di dati campionare e amministrative rappresenta un'opportunità per migliorare la qualità dell'informazione prodotta.

1 L'autore di questo capitolo è Paolo Consolini.

A fronte di un guadagno in termini di qualità del dato, l'approccio in questione richiede, tuttavia, il soddisfacimento di alcuni importanti requisiti, tipici dell'integrazione delle informazioni da molteplici fonti (cfr. van der Laan, 2000). In particolare, occorre preliminarmente garantire che: le unità statistiche siano definite allo stesso modo nelle varie fonti del dato da integrare; tutte le fonti facciano riferimento allo stesso collettivo statistico (copertura); il contenuto informativo delle variabili della fonte "donatrice" sia identico a quello delle variabili "obiettivo" dell'indagine EU-SILC (armonizzazione delle variabili e delle classificazioni). Altro requisito indispensabile per poter applicare con successo tecniche di integrazione di diverse fonti è quello di garantire nel tempo la regolarità del processo di acquisizione dei dati amministrativi e la stabilità del loro contenuto informativo. In Italia, il progetto EU-SILC (che prende il nome di IT-SILC) si innesta in un quadro normativo disciplinato dal Programma statistico nazionale (Psn), coordinato dall'Istat, che autorizza e pianifica il flusso dei dati tra gli Enti fornitori e gli Enti produttori di statistica ufficiale, in un'ottica di condivisione delle esigenze informative degli utenti finali finalizzata all'ampliamento dell'offerta informativa.

8.2 Excursus storico del processo d'integrazione di microdati in EU-SILC

L'integrazione di microdati, sperimentata con successo su alcune componenti di reddito nella prima edizione italiana IT-SILC 2004, è stata perfezionata ed estesa, in modo incrementale, ad altre componenti di reddito nelle successive occasioni d'indagine (Consolini, 2009). Il processo di integrazione, alla cui base si assegnano i profili di reddito a ciascun individuo del campione, è stato migliorato nel tempo col ricorso a nuove fonti amministrative resesi via via disponibili. Ad esempio, grazie allo sfruttamento di altre basi dati INPS e all'inserimento di nuovi campi delle Certificazioni Uniche dell'Agenzia delle Entrate è stato possibile rilevare il collettivo dei percettori di voucher per lavoro occasionale, i lavoratori domestici, i titolari di borse studio, eccetera. A partire dall'edizione 2016 si è avviato un importante processo di riprogettazione del questionario mirato al contenimento del carico statistico (*response burden*). A tal fine, sono stati eliminati tutti i quesiti finalizzati ad acquisire informazioni che, grazie al nuovo processo di acquisizione, esplorazione e validazione delle fonti esterne sviluppato dal team italiano di EU-SILC, erano già reperibili da archivio amministrativo. In totale, la sezione redditi individuali 2016 è stata limitata a 130 quesiti con la soppressione, rispetto alla precedente edizione, di 46 quesiti riferiti a pensioni pubbliche, trasferimenti non pensionistici e saldo fiscale. La riprogettazione è stata resa possibile grazie a numerosi *upgrade* sul fronte dei registri anagrafici della popolazione[2] e all'affinamento di tecniche di generazione dei codici fiscali applicate ai cosiddetti "componenti di fatto"[3] che hanno limitato la perdita di informazione legata ai mancati abbinamenti coi dati amministrativi. Lo storico del processo di validazione sui codici fiscali degli individui inclusi nel campione teorico, trasmessi dall'Istat alla Sogei[4], attesta il netto salto di qualità ottenuto nel corso del tempo: si passa, infatti, dal 92,1% di codici coerenti con l'Anagrafe tributaria per primo anno di rilevazione (EU-SILC 2004) al 99,2% dell'edizione 2016.

2 A decorrere dal maggio 2012 il processo di estrazione dei campioni delle indagini sociali è svolto interamente all'interno dell'Istat tramite l'impiego delle LAC. Queste ultime sono acquisite annualmente dall'Istat presso i servizi demografici dei comuni.

3 Trattasi di coloro che non figurano nella scheda anagrafica del campione teorico, ma che al momento dell'intervista sono riconosciuti dal rilevatore come membri della famiglia. La loro quota si aggira attorno al 2-2,5% degli individui campione.

4 Società di Information and Communication Technology del Ministero dell'Economia e delle Finanze.

L'utilizzo degli arichivi Uniemens dell'INPS nelle ultime due edizioni (EU-SILC 2017 e 2018) ha consentito l'individuazione e correzione degli errori relativi al profilo professionale e al settore di attività economica del lavoratore dipendente riportati nell'intervista diretta. La stessa fonte, come illustrato nel capitolo 10, ha permesso di migliorare la stima sulle integrazioni salariali ai cassintegrati a partire dall'edizione 2016. Inoltre dall'analisi esplorativa del capitolo 11 si evince come la fonte previdenziale giunga a risultati promettenti per la stima dell'indennità di malattia, tali da consigliare il suo impiego nelle prossime edizioni. In futuro, si confida di avviare nuove sperimentazioni sulle basi dati Uniemens per gettare luce su prestazioni sociali sinora inesplorate, quali ad esempio la maternità e la cura dei figli.

8.3 Presentazione dei contributi della parte terza

In questa parte terza verranno presentati altri quattro contributi. Il primo riporta un'analisi dell'impatto dell'informazione di fonte amministrativa sulle stime finali di talune componenti di reddito, con relativa valutazione della possibilità di ottenere buone stime dall'uso esclusivo di questa fonte. Il secondo e il terzo forniscono degli esempi di come sia possibile migliorare la stima di alcune componenti di reddito (integrazione salariale e indennità di malattia) tramite lo sfruttamento di nuovi giacimenti informativi di natura amministrativa (Uniemens e altre fonti INPS). Infine, l'ultimo contributo si pone l'obiettivo di valutare la presenza di distorsioni nelle stime dell'indagine IT-SILC relative ai redditi e ai principali indicatori distributivi (rischio di povertà e indicatori di disuguaglianza), connesse ad un'imperfetta rappresentazione della "vera" struttura reddituale della popolazione. Nella fattispecie, si è fatto riscorso alla banca dati reddituale del MEF (BDR) che raccoglie le principali informazioni fiscali sulle dichiarazioni dei redditi per il complesso dei contribuenti residenti in Italia.

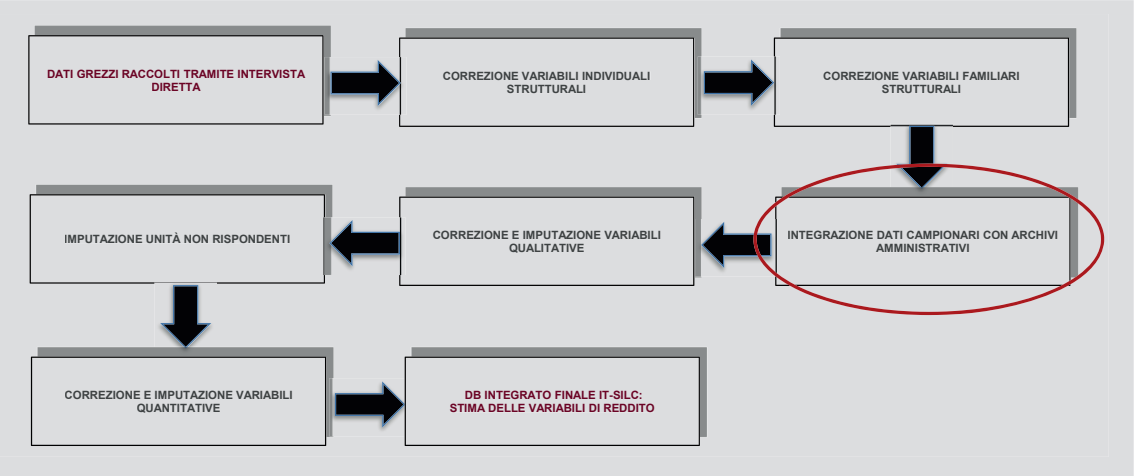
9. L'IMPATTO DEI DATI AMMINISTRATIVI SULLE STIME FINALI DEI REDDITI[1]

9.1 Introduzione

Questo lavoro, presentato al "Workshop on best practices for EU-SILC revision" (Londra, 16-17 settembre 2015) e alla "LIII Riunione Scientifica SIEDS" (Roma 26-28 maggio 2016), illustra i principali risultati delle analisi volte a valutare l'impatto dei dati amministrativi sulle stime finali delle variabili di reddito dell'indagine IT-SILC.

Per meglio comprendere il contesto dove si colloca l'utilizzo di dati di fonte amministrativa nel processo di produzione di IT-SILC si può prendere a riferimento il diagramma di flusso in Figura 9.1. Si noti che l'integrazione, a livello micro, tra dati di fonte campionaria e dati di fonte amministrativa si colloca nella parte centrale del processo, preceduta dalla correzione delle variabili strutturali individuali e familiari, e seguita dalla correzione e dall'imputazione delle variabili individuali qualitative e quantitative. L'integrazione tra dati campionari e dati amministrativi può essere vista, più in generale, come un unico processo costituito da una serie di fasi che partono dall'analisi delle fonti disponibili per i diversi fenomeni oggetto di studio e si concludono con la riconciliazione delle situazioni di incongruenza tra tipi di percettore e/o tra valori di reddito nelle varie fonti. Brevemente, il metodo utilizzato in IT-SILC per integrare i micro dati provenienti dalle diverse fonti è il record linkage deterministico, con codice fiscale come chiave individuale di linkage. Aspetti cruciali del processo di integrazione sono la riclassificazione e la riconciliazione delle componenti di reddito nel database linkato: in particolare, l'identificazione dei percettori per le diverse componenti di reddito e la formulazione di un sistema di ipotesi per riconciliare le situazioni di incongruenza tra tipi di percettore e/o tra valori di reddito (Consolini, 2009).

Figura 9.1 - Principali fasi del processo di produzione di IT-SILC



1 I paragrafi 9.1, 9.2 e 9.3 sono stati redatti da Francesca Lariccia; i paragrafi 9.4 e 9.5 sono stati redatti da Clodia Delle Fratte.

Obiettivo del presente lavoro è valutare l'impatto dei dati amministrativi sulle stime delle variabili di reddito dell'indagine IT-SILC. Più in particolare, il lavoro si propone di:

- mostrare come le stime finali delle variabili di reddito ottenute utilizzando il database Integrato Finale IT-SILC siano di qualità migliore rispetto alle stime che si otterrebbero considerando solo i dati di fonte campionaria;
- analizzare le stime finali delle variabili di reddito che si otterrebbero utilizzando solo i dati di fonti amministrativa;
- valutare per quali componenti di reddito è possibile produrre buone stime utilizzando esclusivamente i dati di fonte amministrativa, ossia sostituendoli a quelli di fonte campionaria ed eliminando quindi alcune parti dell'intervista diretta.

9.2 Dati e metodi

Per raggiungere i precedenti obiettivi, a partire dal database Integrato Finale IT-SILC (edizione 2011), sono stati costruiti il database Solo Campione (con solo i dati di fonte campionaria) e il database Solo Fisco (con solo i dati di fonte amministrativa). Questa fase è risultata la più onerosa del lavoro. Sono state quindi confrontate le stime finali delle variabili di reddito ottenute, focalizzandosi in particolare su:

- distribuzione del reddito totale familiare netto;
- rischio di povertà;
- profili di reddito delle principali componenti del reddito individuale, ossia reddito da lavoro dipendente, da lavoro autonomo e da pensione. Queste tre componenti, infatti, rappresentano le principali fonti di reddito delle famiglie e, inoltre, sono le componenti sulle quali IT-SILC ha investito maggiormente in termini di acquisizione e armonizzazione di informazioni dagli archivi amministrativi.

I confronti sono stati effettuati prima tra database Solo Campione e database Integrato Finale IT-SILC e, successivamente, tra database Solo Fisco e Integrato Finale IT-SILC.

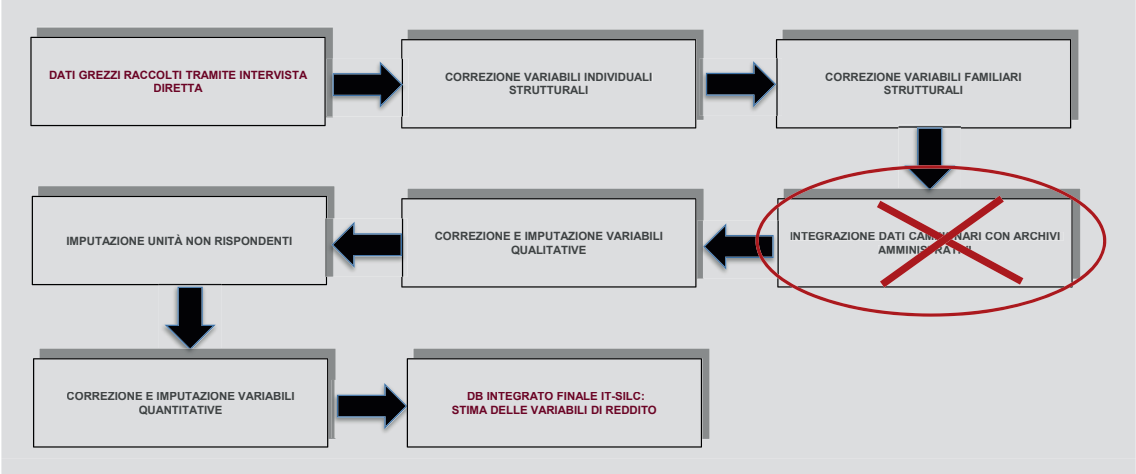
9.3 Risultati: database IT-SILC Solo Campione vs Integrato Finale

In questa sezione vengono presentati la costruzione del database Solo Campione e i principali risultati ottenuti dal confronto tra le stime di reddito ottenute da questo database e quelle ottenute dal database Integrato Finale IT-SILC.

9.3.1 Costruzione del database Solo Campione

Per stimare le variabili di reddito, considerando solo le informazioni di fonte campionaria, a partire dal database dei dati grezzi raccolti sul campione del 2011 tramite intervista diretta con tecnica CAPI, sono state applicate tutte le fasi del processo di produzione di IT-SILC, eccetto l'integrazione con gli archivi amministrativi (Figura 9.2). L'output finale è il database Solo Campione che ha ovviamente stessa numerosità campionaria del database Integrato Finale IT-SILC (n=47.841).

Figura 9.2 - Costruzione del database Solo Campione: stima delle variabili di reddito



9.3.2 Distribuzione del reddito totale familiare netto

Il confronto tra le distribuzioni del reddito totale familiare netto[2] nei due database ha evidenziato che il database Integrato Finale IT-SILC stima redditi più elevati rispetto al database Solo Campione (Delle Fratte e Lariccia, 2015 e 2016): il processo di integrazione con i dati di fonte amministrativa, infatti, consente di recuperare informazioni sui redditi minori o secondari spesso omessi durante l'intervista diretta perché dimenticati o ritenuti irrilevanti (effetto memoria). Coerentemente con le differenze appena descritte, l'ordinamento[3] delle famiglie secondo il reddito totale familiare netto non coincide nei due database, come si evince dalla Tavola 9.1 che mostra la distribuzione delle famiglie per quinto di appartenenza nel database Solo Campione (prima colonna) e nel database Integrato Finale IT-SILC (seconda colonna). Solo il 43,3% delle famiglie si colloca nello stesso quinto di reddito nei due database, come si nota osservando i valori evidenziati in rosso nella Tavola 9.1a. Se si analizzano, invece, le famiglie che non restano nella stessa posizione, ossia che cambiano quinto di reddito, emerge che questo cambiamento è più spesso verso un quinto più elevato: ben l'8,3% delle famiglie appartiene nel Solo Campione alla parte più povera della distribuzione, ossia al primo o secondo quinto, mentre in IT-SILC ai quinti più ricchi, ossia il quarto o il quinto (valori in verde nella Tavola 9.1b); viceversa, solo l'1,6% delle famiglie passa dal segmento superiore a quello inferiore della distribuzione dei redditi (valori in azzurro nella Tavola 9.1b).

Allo scopo di valutare la maggiore accuratezza della classificazione ottenuta con IT-SILC, è stata calcolata la grave deprivazione materiale[4]. Le famiglie che si trovano in condizioni di grave deprivazione, infatti, sono sempre le stesse, a prescindere dal database considerato, poiché le informazioni che definiscono tale condizione sono desunte dall'intervista campionaria e, non essendo variabili di tipo reddituale, non subiscono alcuna modifica durante il processo di integrazione con i dati di fonte amministrativa.

2 Il reddito totale familiare netto considerato in questo studio è dato dal reddito totale disponibile delle famiglie inclusi i fitti figurativi ed esclusi i rimborsi di interesse a mutuo e i fringe benefits.

3 In ognuno dei due database le famiglie sono ordinate secondo il reddito netto familiare equivalente con fitti imputati. I quinti sono quindi calcolati in modo endogeno.

4 È un indicatore Europa 2020. È definito come la percentuale di persone in famiglie che registrano almeno quattro segnali di deprivazione materiale sui nove stabiliti a livello europeo.

Le famiglie che in IT-SILC si collocano in un quinto più elevato della distribuzione dei redditi si trovano più raramente in condizioni di grave deprivazione (8,9%) delle famiglie che si collocano nei quinti più poveri secondo entrambi i database (21,8%). Si può concludere che le stime dei redditi familiari ottenute utilizzando il database Integrato Finale IT-SILC sono più coerenti con le condizioni materiali di vita delle famiglie rispetto alle stime ottenute considerando solo i dati di fonte campionaria. Questo risultato conferma un miglioramento nell'accuratezza delle stime.

Tabella 9.1 - Famiglie per quinto di appartenenza nel database Solo Campione e nel database Integrato Finale IT-SILC. Anno 2011 (valori percentuali)

1a

DB SOLO CAMPIONE	DB INTEGRATO FINALE	%
Quinti	Quinti	
	1	**14,2**
	2	
1	3	
	4	
	5	
	1	
	2	**7,3**
2	3	
	4	
	5	
	1	
	2	
3	**3**	**5,7**
	4	
	5	
	1	
	2	
4	3	
	4	**6,1**
	5	
	1	
	2	
5	3	
	4	
	5	**10,0**

1b

DB SOLO CAMPIONE	DB INTEGRATO FINALE	%
Quinti	Quinti	
	1	14,2
	2	7,1
1	3	3,2
	4	**1,7**
	5	**1,0**
	1	4,1
	2	7,3
2	3	6,1
	4	**4,0**
	5	**1,6**
	1	1,2
	2	4,2
3	3	5,7
	4	4,5
	5	3,1
	1	**0,2**
	2	**1,2**
4	3	4,0
	4	6,1
	5	4,7
	1	**0,0**
	2	**0,2**
5	3	1,0
	4	3,6
	5	10,0
Totale		100,0

9.3.3 Rischio di povertà

Si confrontano ora le stime dell'indicatore Europa 2020 "rischio di povertà" seguendo lo stesso approccio utilizzato per l'analisi della distribuzione dei redditi. Le persone che vivono in famiglie a rischio di povertà sono il 21,2% secondo il database Solo Campione, il 19,6% secondo il database Integrato Finale IT-SILC. L'86,8% degli individui è classificato in modo omogeneo nei due database (valori in rosso nella Tavola 9.2).

Tavola 9.2 - Individui per condizione di rischio di povertà nel database Solo Campione e nel database Integrato Finale IT-SILC. Anno 2011 (valori percentuali)

DB SOLO CAMPIONE	DB INTEGRATO FINALE	%
A rischio di povertà	A rischio di povertà	
Si	**Si**	**13,8**
Si	No	7,4
No	Si	5,8
No	**No**	**73,0**
		Totale 100,0

Le persone che sono classificate in famiglie a rischio di povertà soltanto secondo il database Solo Campione sono più raramente in condizioni di grave deprivazione materiale (21,9%) rispetto a quelle che risultano a rischio di povertà in entrambi i database (31,1%) come si osserva nella Tavola 9.3. Anche questo risultato conferma che le stime ottenute tramite l'utilizzo del database Integrato Finale IT-SILC sono migliori, in termini di accuratezza, rispetto a quelle che si otterrebbero con l'uso esclusivo di informazioni di fonte campionaria.

Tavola 9.3 - Individui per condizione di rischio di povertà nel database Solo Campione e nel database Integrato Finale IT-SILC, e Grave deprivazione materiale. Anno 2011 (valori percentuali)

DB SOLO CAMPIONE	DB INTEGRATO FINALE	%	Grave deprivazione materiale
Rischio di poverà	Rischio di poverà		
Si	**Si**	**13,8**	**31,1**
Si	**No**	**7,4**	**21,9**
No	Si	5,8	15,7
No	No	73,0	5,9
		Totale 100,0	

9.3.4 Principali componenti del reddito individuale e loro profili

I risultati illustrati nei lavori precedenti (Delle Fratte e Lariccia, 2015 e 2016) hanno mostrato che il processo di integrazione tra informazioni di fonte campionaria e di fonte amministrativa ha un impatto importante anche sui livelli e la distribuzione delle singole componenti del reddito individuale, nonché sul numero di individui che percepiscono le singole tipologie reddituali (percettori): per tutte le componenti di reddito considerate - reddito da lavoro dipendente, da lavoro autonomo e da pensione - il database Integrato Finale IT-SILC stima un maggior numero di percettori e un reddito medio più elevato di quanto stimato utilizzando solo i dati campionari; tale incremento è dovuto alle informazioni omesse durante l'intervista e recuperate dai dati amministrativi.

L'impatto descritto si ripercuote anche sulla composizione dei redditi individuali percepiti (profili di reddito): per confrontare quali tipologie di reddito possiedono gli individui, in entrambi i database è stata costruita una variabile "profilo di reddito" che riporta, per ogni persona, la presenza/assenza di redditi da lavoro dipendente, da lavoro autonomo, da pensione; la modalità "nessuna componente" indica che l'individuo non percepisce nessuna delle tre tipologie di reddito considerate, non che non possiede in assoluto alcun reddito.

L'86,2% degli individui presenta lo stesso profilo di reddito secondo il database Solo Campione e secondo il database Integrato Finale IT-SILC (Tavola 9.4). Il 12,4%, invece, a seguito del processo di integrazione con dati di fonte amministrativa, risulta avere più tipologie di reddito rispetto a quanto risulta dall'intervista diretta, ossia ha un profilo più articolato: questo è attribuibile a tutte quelle situazioni in cui l'intervistato non ricorda, omette, classifica erroneamente[5], non vuole o non sa rispondere (effetto memoria, redditi secondari o frazionati, *under-reporting*, interviste *proxy*). Viceversa, avere meno tipi di reddito rispetto a quanto rilevato nell'intervista diretta è una circostanza del tutto trascu-

5 Errori di classificazione da parte del rispondente possono comportare l'aggiunta di una componente di reddito. Per esempio intervistati che dichiarano reddito da capitale invece sono autonomi, o che dichiarano indennità di disoccupazione invece sono dipendenti o pensionati.

rabile e riguarda solo lo 0,6% dei casi. Infine, Il restante 0,8% sono casi di misclassificazione (stesso numero di redditi ma classificati diversamente).

La situazione che si verifica più frequentemente (Tavola 9.4) è quella in cui dai dati campionari la persona risulta senza nessuna delle tre componenti di reddito, mentre, nel database Integrato Finale IT-SILC ha almeno una tipologia reddituale (situazione che riguarda il 23,1% di chi nell'intervista diretta non dichiara alcun reddito). Gli individui che si trovano in questa casistica sono soprattutto con un'età superiore a 60 anni, uomini.

Analogamente, il 12,9% di chi durante l'intervista dichiara di avere solo redditi da lavoro autonomo, nel database Integrato Finale IT-SILC ha anche altri redditi. Si tratta prevalentemente di persone di oltre 60 anni o giovani sotto ai 30 e quindi di redditi frazionati, secondari o da lavori minori, nel caso degli anziani anche da redditi da pensione.

Infine, il 9,5% di chi, dal database Solo Campione, ha solo redditi da pensione, nel database Integrato Finale IT-SILC risulta avere anche altre componenti di reddito: in questo caso si tratta per lo più di uomini, under75 (soprattutto 45-59enni), che oltre alla pensione hanno redditi minori da lavoro.

Tabella 9.4 - Principali componenti di reddito: individui per profili di reddito nel database Solo Campione e nel database Integrato Finale IT-SILC. Anno 2011 (valori per 100 interviste individuali)

DB SOLO CAMPIONE	DB INTEGRATO FINALE	%
Nessuna delle tre componenti di reddito	**Nessuna delle tre componenti di reddito**	**76,9**
	Almeno una componente di reddito	23,1
Solo reddito da lavoro dipendente	**Solo reddito da lavoro dipendente**	**90,4**
	Reddito da lavoro dipendente + reddito da lavoro autonomo e/o pensione	9,0
	Nessuna delle tre componenti di reddito	0,6
Solo reddito da lavoro autonomo	**Solo reddito da lavoro autonomo**	**86,5**
	Reddito da lavoro autonomo + reddito da lavoro dipendente e/o pensione	12,9
	Nessuna delle tre componenti di reddito	0,6
Solo reddito da pensione	**Solo reddito da pensione**	**90,3**
	Reddito da pensione + reddito da lavoro dipendente e/o autonomo	9,5
	Nessuna delle tre componenti di reddito	0,2

9.4 Risultati: database IT-SILC Solo Fisco vs Integrato Finale

Come precedentemente studiato per il database Solo Campione, in questo paragrafo vengono presentati sia la costruzione del database Solo Fisco, sia i principali risultati emersi dal confronto tra le stime di reddito ottenute da questo database e quelle del database Integrato Finale IT-SILC.

9.4.1 Costruzione del database Solo Fisco

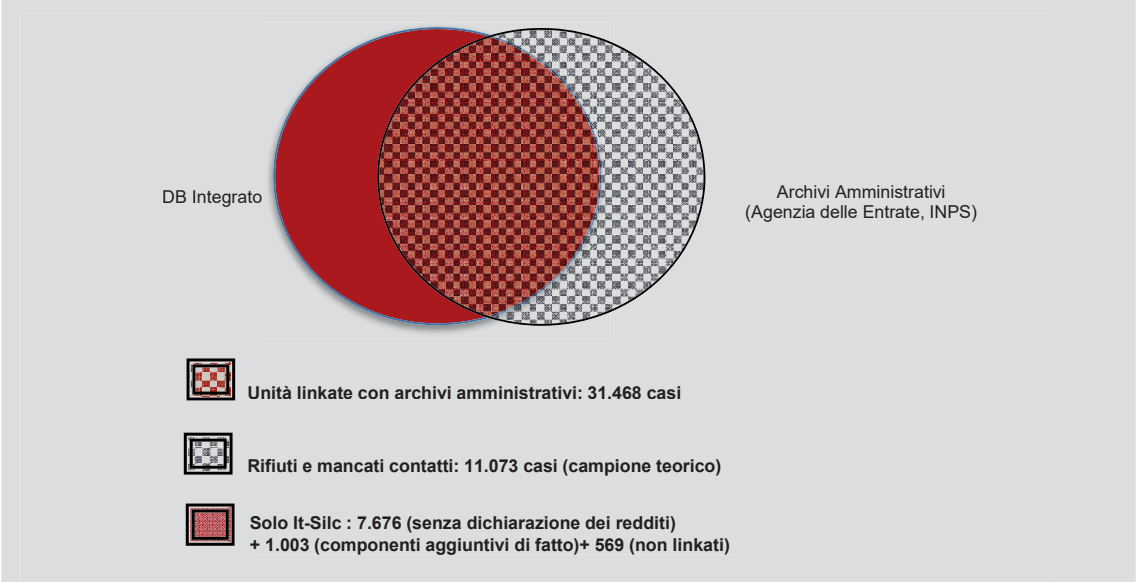
Per ottenere le stime dei redditi basate solo su dati di fonte fiscale, agli individui campione del 2011 sono state agganciate le variabili di reddito del database Db_integrato_pl, costituito da informazioni relative a pensioni (fonte INPS) e dichiarazioni dei redditi (fonte Agenzia delle Entrate) (Figura 9.3). Data la disponibilità, nell'ambito dell'indagine IT-SILC, dei codici fiscali delle persone fisiche tale aggancio è stato realizzato tramite record linkage deterministico.

Le unità linkate sono pari a 31.468, i restanti 9.248[6] individui-campione non trovano aggancio perché:

- sono componenti aggiuntivi di fatto, ossia individui intervistati che non sono inclusi nel campione teorico (*sample frame*), ma entrano a far parte del campione al momento dell'intervista (*effective sample*)[7];
- sono affetti da errore nel codice fiscale;
- sono lavoratori in nero (*grey economy*);
- sono individui che non hanno copertura nelle fonti amministrative, come per esempio chi ha solo redditi da capitale, ossia che vive solo di rendita.

Il recupero di tutti questi casi è avvenuto con strategie diverse. Per quanto riguarda i componenti aggiuntivi di fatto (1.003 casi) e i mancati agganci per errori nel codice fiscale (569 casi), tra le possibili soluzioni, si è scelto di imputate le variabili di reddito. Più in particolare, mantenendo quanto dichiarato nel prospetto redditi durante l'intervista, ossia la percezione o meno un determinato reddito, tramite una tecnica basata sulla costruzione di modelli di regressioni sequenziali multiple implementata nel software IVEware, sono state imputate le variabili di reddito usando come bacino di donatori i dati fiscali presenti nel Db_integrato_pl. Infine, quei casi che non trovano copertura nelle fonti amministrative (7.676 casi), non per errori nel codice fiscale, che risulta comunque validato e corretto, ma perché effettivamente non hanno dichiarato alcun reddito di quelli considerati in IT-SILC, sono stati inclusi nel database Solo Fisco con reddito pari a zero.

Figura 9.3 - Costruzione del database Solo Fisco: stima delle variabili di reddito



9.4.2 Principali componenti del reddito individuale e loro profili

Dalle analisi presentate nei lavori precedenti (Delle Fratte e Lariccia, 2015 e 2016) è emerso anche come, paragonando il database Solo Fisco con il database Integrato Finale

6 Il database Solo Fisco è composto da 40.716 casi poiché include solo individui con età maggiore di 15 anni.

7 Si sottolinea che, per motivi organizzativi indotti dalla tempistica, la disponibilità dei codici fiscali è limitata agli individui di partenza in ogni edizione dell'indagine (campione teorico): si tratta di individui estratti dalle liste anagrafiche per la prima *wave*, o di individui in liste esito delle interviste precedenti per le *wave* dalla seconda in poi.

IT-SILC, quest'ultimo stimi un maggior numero di percettori per tutte e tre le componenti di reddito esaminate. Nel caso di redditi da lavoro si tratta di lavoratori del sommerso, infatti il rispondente durante l'intervista diretta dichiara di avere redditi che invece non risultano negli archivi amministrativi in quanto non denunciati: ciò riguarda i redditi da lavoro dipendente, (+8,1%), ma soprattutto i redditi da lavoro autonomo (+ 20,2%), dove i rapporti di carattere lavorativo svolti, anche solo in parte, in violazione delle vigenti normative di carattere tributario e contributivo sono più frequenti. Contrariamente ai redditi da lavoro, le pensioni non sono affette da evasione fiscale e il lieve aumento dei percettori da pensione (2,6%) è spiegato dalla mancata copertura delle fonti amministrative. I lavori precedenti mostrano anche che il reddito medio da lavoro autonomo stimato dal database integrato è più elevato di quello stimato utilizzando solo i dati di fonte amministrativa (+19,4%): tale incremento è attribuibile alla cosiddetta "regola del valore massimo"[8] adottata durante il processo di integrazione.

L'impatto descritto si ripercuote anche sulla composizione dei redditi individuali percepiti (profili di reddito), come già emerso dal confronto tra database Solo Campione e database Integrato Finale IT-SILC. Analogamente, sono stati paragonati i profili di reddito ottenuti utilizzando solo le informazioni amministrative con quelli ottenuti a seguito del processo di integrazione. Il 93,0% degli individui presenta lo stesso profilo di reddito secondo il database Solo Fisco e secondo il database Integrato Finale IT-SILC: la percentuale di profili coincidenti è più alta di quella emersa dal confronto con il database Solo Campione ed è così elevata per costruzione, infatti nel processo di integrazione il profilo proveniente da fonte amministrativa "comanda" nel caso di incongruenza.

La situazione che si verifica più frequentemente (Tavola 9.5) è quella in cui dalle fonti amministrative la persona risulta senza nessuna delle tre componenti di reddito, mentre, nel database Integrato Finale IT-SILC ha almeno una tipologia reddituale (situazione che riguarda il 25,3% di chi dagli archivi non risulta avere nessuna delle tre tipologie di reddito considerate). Gli individui che si trovano in questa casistica sono soprattutto di età compresa tra 30 e 59 anni, uomini. Si tratta verosimilmente di componenti di reddito in nero.

Tabella 9.5 - Principali componenti di reddito: Individui di 15 anni o più per profili di reddito nel database Solo Fisco e nel database Integrato Finale IT-SILC. Anno 2011 (valori per 100 interviste individuali)

DB SOLO FISCO	DB INTEGRATO FINALE	%
Nessuna delle tre componenti di reddito	**Nessuna delle tre componenti di reddito**	**74,7**
	Almeno una componente di reddito	25,3
Solo reddito da lavoro dipendente	**Solo reddito da lavoro dipendente**	**95,8**
	Reddito da lavoro dipendente + reddito da lavoro autonomo e/o pensione	4,1
	Nessuna delle tre componenti di reddito	0,1
Solo reddito da lavoro autonomo	**Solo reddito da lavoro autonomo**	**92,0**
	Reddito da lavoro autonomo + reddito da lavoro dipendente e/o pensione	8,0
	Nessuna delle tre componenti di reddito	0,0
Solo reddito da pensione	**Solo reddito da pensione**	**98,0**
	Reddito da pensione + reddito da lavoro dipendente e/o autonomo	2,0
	Nessuna delle tre componenti di reddito	0,0

Il 95,8% di coloro che hanno solo un reddito da lavoro dipendente nel database Solo Fisco si trova nella stessa condizione nel database Integrato Finale IT-SILC; il 4,1%, invece,

8 Secondo la "regola del valore massimo" si attribuisce il massimo tra i valori rilevati nell'indagine campionaria e nelle fonti fiscali. Si assume, infatti, che sia le informazioni provenienti dall'intervista diretta, che quelle desunte dagli archivi amministrativi sottostimino il valore "vero" del reddito autonomo e che, pertanto, usando il massimo tra i valori rilevati tra le due fonti, l'errore di misura sia minimizzato.

percepisce anche un altro tipo di reddito, nella metà dei casi si tratta di un reddito da lavoro autonomo. Sono soprattutto uomini, tra 64 e 74 anni, residenti nel Mezzogiorno. Analogamente, l'8,0% di quanti dai dati fiscali risultano essere percettori solo di reddito autonomo, nel database Integrato Finale IT-SILC ha anche altri redditi. Infine, tra i percettori di sole pensioni secondo il database Solo Fisco ben il 98,0% percepisce solo questo tipo di reddito anche dopo l'integrazione con i dati provenienti da intervista diretta. Il restante 2,0% percepisce anche altri redditi.

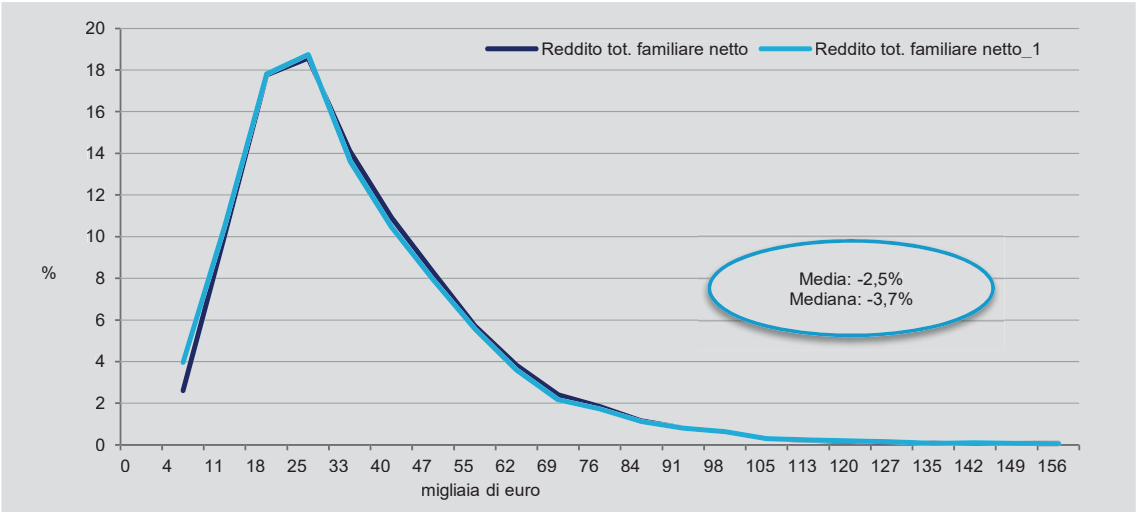
9.4.3 Quali delle componenti di reddito di fonte campionaria potrebbero essere sostituite con dati di fonte amministrativa?

Dopo aver mostrato che l'integrazione con informazioni amministrative comporta un miglioramento nella qualità delle stime finali delle variabili di reddito rispetto all'utilizzo dei soli dati campionari e dopo aver analizzato le stime che si otterrebbero utilizzando solo i dati di fonti amministrativa, ultimo obiettivo del presente lavoro è valutare per quali componenti di reddito è possibile produrre buone stime utilizzando esclusivamente i dati di fonte amministrativa, ossia sostituendo l'intervista diretta. Per poter valutare tale impatto è stato confrontato il reddito totale familiare netto del database Integrato Finale IT-SILC con il reddito totale familiare netto che si otterrebbe sostituendo una alla volta le tre principali componenti di reddito (da lavoro dipendente, da lavoro autonomo e da pensione) con l'informazione analoga ma proveniente da fonte amministrativa.

Nel Grafico 9.1 l'impatto è valutato rispetto ai soli redditi da lavoro dipendente. Sostituendo i dati di fonte amministrativa di questa componente di reddito ("reddito totale familiare netto_1" in figura) si ottiene un valore medio e mediano di reddito familiare inferiore (rispettivamente -2, 5% e -3,7%) anche se le due distribuzioni sono simili sia nella forma che nei livelli.

Come già descritto nei risultati precedenti, il reddito da lavoro autonomo è la componente per la quale emergono le differenze più consistenti (Grafico 9.2). Le due curve differiscono maggiormente in livelli e distribuzioni, in particolare si nota come nel reddito familiare costruito utilizzando solo i redditi autonomi di origine amministrativa ("reddito totale familiare netto_2" in figura) i redditi bassi pesano di più con un valore medio e mediano inferiore di 4,5% e 4,2% rispetto al reddito totale familiare netto del database Integrato Finale IT-SILC.

Grafico 9.1 - Distribuzione del reddito totale familiare netto finale e con reddito da lavoro dipendente solo da fonte amministrativa (reddito totale familiare netto_1). Anno 2011



Infine, sostituendo nella costruzione del reddito totale familiare netto solo il reddito da pensione di fonte amministrativa ("reddito totale familiare netto_3" nel Grafico 9.3), si nota come le due curve di reddito siano perfettamente sovrapposte, infatti il reddito totale familiare si discosta solo di 0,6% per la media e di -0,1% per la mediana da quello finale. Tali dati confermano quanto già precedentemente detto sui redditi da pensione.

Grafico 9.2 - Distribuzione del reddito totale familiare netto finale e con reddito da lavoro autonomo solo da fonte amministrativa (reddito totale familiare netto_2). Anno 2011

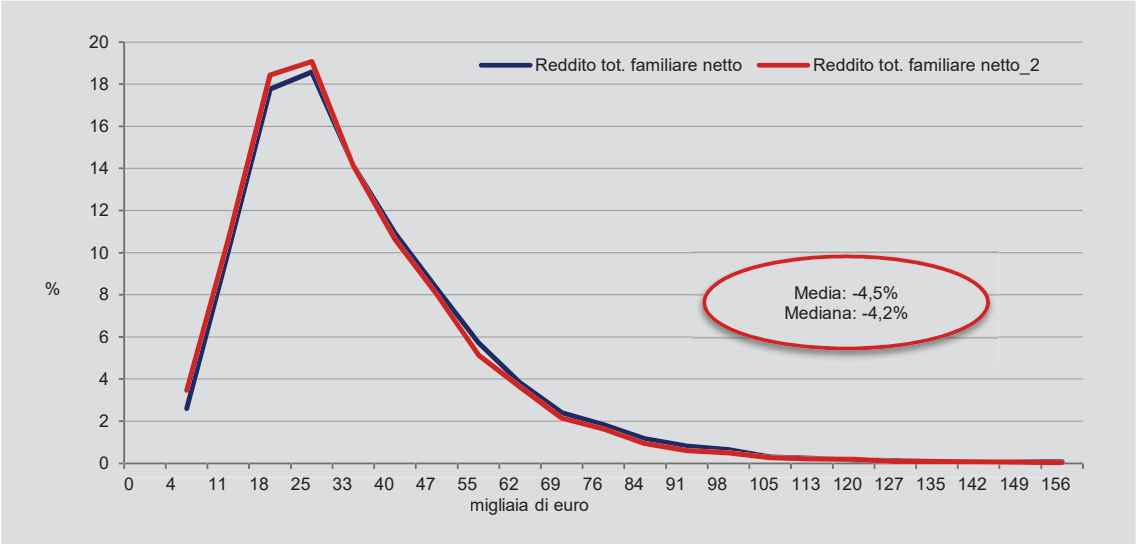
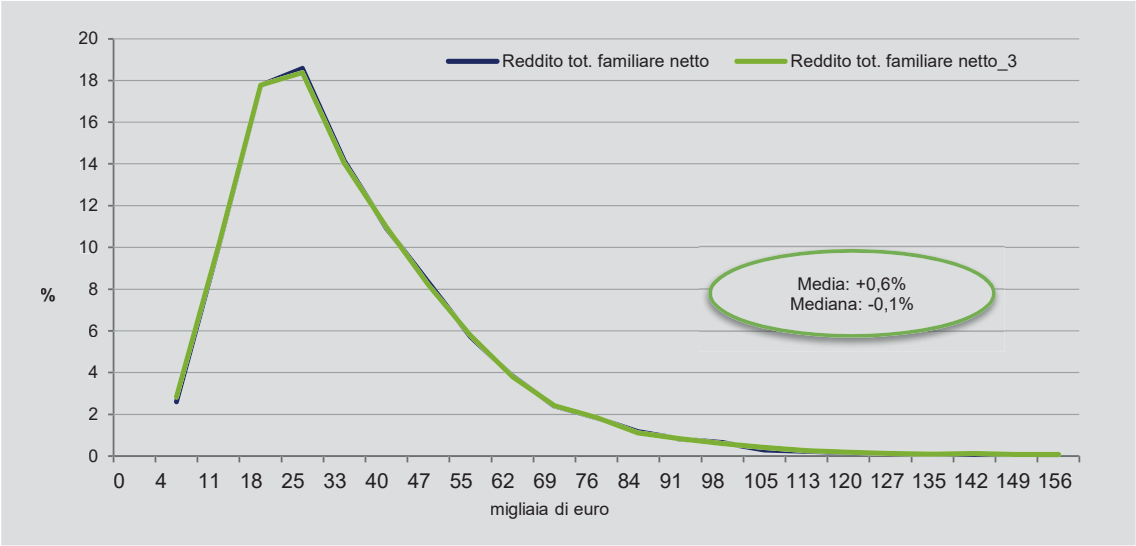


Grafico 9.3 - Distribuzione del reddito totale familiare netto finale e con reddito da pensione solo da fonte amministrativa (reddito totale familiare netto_3). Anno 2011



9.5 Conclusioni

Per concludere, il lavoro mostra che le stime finali delle variabili di reddito ottenute utilizzando il database Integrato Finale IT-SILC sono di qualità migliore rispetto a quelle ottenute considerando il database Solo Campione. In termini di accuratezza:

- riduzione della mancata risposta selettiva e degli errori di misura (effetto memoria/ *telescoping effect, under-reporting*);
- maggior corrispondenza tra le condizioni monetarie (livelli di reddito) e le condizioni materiali (grave deprivazione materiale).

E in termini di completezza: maggiore copertura rispetto alla popolazione di riferimento, in particolare per i redditi da pensione.

Nonostante i vantaggi sopra menzionati, l'utilizzo del database Integrato Finale IT-SILC, rispetto al database Solo Campione, comporta alcune criticità relative a:

- tempestività: in particolare, vincoli legati alla tempistica di acquisizione degli archivi amministrativi;
- coerenza: i dati amministrativi sono raccolti con obiettivi e definizioni proprie, non sempre coincidenti con quelle statistiche;
- efficienza: maggior "pesantezza" del processo di produzione delle stime finali dovuto alla fase di integrazione a livello micro tra dati campionari e dati amministrativi.

Inoltre, le analisi mostrano che per alcune componenti di reddito - redditi da pensione e, in parte, redditi da lavoro dipendente - è possibile produrre delle buone stime usando esclusivamente i dati di fonte amministrativa, ossia sostituendoli a quelli di fonte campionaria ed eliminando quindi alcune parti dell'intervista diretta. Questo ha conseguenze positive rispetto a:

- costi di gestione dell'indagine e carico sui rispondenti: riduzione dovuta all'eliminazione di quei quesiti del questionario non più necessari poiché l'informazione è desunta da fonte amministrativa;
- efficienza: processo più snello dovuto all'eliminazione dell'integrazione per quelle componenti di reddito per le quali si usa esclusivamente l'informazione di fonte amministrativa.

Tale processo di parziale sostituzione delle informazioni da questionario con quelle amministrative presenta gli stessi svantaggi in termini di tempestività e coerenza sopra menzionati. In questo caso, però, tali criticità pesano maggiormente dato il ruolo esclusivo degli archivi.

10. LA STIMA DELLA CASSA INTEGRAZIONE ANTICIPATA DAI DATORI DI LAVORO[1]

10.1 Introduzione

L'obiettivo informativo che ci si è prefissati nell'implementazione del nuovo processo di stima dei redditi è di scorporare dalla retribuzione la cassa integrazione erogata in busta paga, ovvero quella componente anticipata dal datore per conto dell'Inps, che andrebbe a rigore classificata tra le prestazioni sociali a copertura del rischio di disoccupazione. Le fonti fiscali impiegate nel processo di stima di EU-SILC consentono, infatti, di calcolare il reddito annuo (al lordo e al netto della tassazione) versato al dipendente e al tempo stesso di ricavare le mensilità retribuite, ma non permettono di suddividere ulteriormente le voci economiche presenti in busta paga secondo livelli di classificazione più fini. In base alle linee guida definite nel documento metodologico EU-SILC la definizione di reddito da lavoro dipendente (PY010N/G) deve rispondere al seguente principio:

> *"Gross employee cash or near cash income (PY010N-G) includes the followings items: … Payments made by employers to an employee in lieu of wages and salaries through a social insurance scheme when unable to work through sickness, disability or maternity leave where such payment cannot be separately and clearly identified as social benefits … in case these payment can be identified they should be included in appropriate benefits variables instead;"*

Si ammette, dunque, che nella variabile target reddito la lavoro dipendente possano confluire componenti spurie (prestazioni sociali) qualora, a causa di deficit informativi, risulti impossibile separarle dal resto della retribuzione. L'obiettivo del presente lavoro è di colmare parte delle lacune, partendo dal *trattamento* della cassa integrazione che confluisce nella target "unemployment benefits"; il quale in passato, a causa di carenze informative, veniva convenzionalmente assimilato alla retribuzione da lavoro dipendente. Tale innovazione costituisce una base per l'affinamento del calcolo delle diverse componenti di reddito secondo la classificazione armonizzata a livello europeo.

Attualmente le fonti amministrative a disposizione dell'indagine italiana IT-SILC consentono la copertura della sola cassa integrazione erogata direttamente dall'INPS[2] (ordinaria, straordinaria, agricola e fondo speciale trasporto aereo) che costituisce tuttavia una quota minoritaria della spesa previdenziale complessiva per questo trattamento[3]. Infatti, la parte preponderante è rappresentata dalla cassa integrazione (indiretta) anticipata dal datore di lavoro per conto dell'INPS.

---

1 Il capitolo è opera di Maria Cirelli e Paolo Consolini, i paragrafi 10.1, 10.5 e 10.6 sono da attribuire a Paolo Consolini; i restanti paragrafi 10.2, 10.3 e 10.4 a Maria Cirelli.
Un particolare ringraziamento va rivolto alla collega Silvia Pacini del Servizio SWB dell'Istat, esperta di fonti amministrative e responsabile delle iniziative su Registro del lavoro e Registro RACLI, nonché alle rispettive collaboratrici Francesca Rossetti (da poco trasferitasi alla Raccolta Dati) e Sara Gigante, per i suggerimenti metodologici. Chiaramente eventuali errori, omissioni e imprecisioni nell'applicazione della strategia finale sono da imputare esclusivamente agli autori.

2 L'impresa che versa in una situazione di grave difficoltà finanziaria, tale da impedirle il pagamento degli stipendi, può chiedere tramite il Ministero del lavoro che la CIG sia erogata direttamente dall'INPS.

3 Per una disamina più approfondita dell'argomento, sebbene non molto aggiornata, si rimanda al contributo di Consolini e De Carli, 2002.

All'inizio nell'indagine IT-SILC 2005, della cassa integrazione si rilevava soltanto il numero dei mesi e l'importo mensile netto e quest' ultimo veniva fatto confluire convenzionalmente nel reddito da lavoro dipendente. A partire dall'edizione di IT-SILC 2012 fino al 2015, sono stati introdotti i quesiti che distinguono quando la cassa integrazione viene erogata direttamente dall'INPS o se invece risulta anticipata dal datore di lavoro in busta paga. In tal modo, nell'indagine si trovano una serie di quesiti riguardanti: il numero dei mesi in cui percepiva il trattamento di cassa integrazione ordinaria (Cigo), straordinaria (Cigs), in deroga o agricola, la percentuale di riduzione del suo orario di lavoro e la forma di pagamento (erogazione in busta paga o versamento diretto dall'INPS). Il calcolo dell'importo della CIG in busta paga si ricava attraverso le risposte ai quesiti di cui sopra in funzione del dato retributivo. In questo capitolo, verrà descritta invece la nuova metodologia introdotta a partire dal 2016 che consente di stimare in modo più accurato l'integrazione salariale anticipata sfruttando una serie di informazioni presenti nella fonte Uniemens dell'INPS.

10.2 Le tipologie di cassa integrazione: caratteristiche e riferimenti normativi

La cassa integrazione guadagni rappresenta una prestazione sociale in denaro erogata dall'INPS, la cui istituzione risale al decreto legislativo n.788 del 1945. Il trattamento permette alle aziende di superare periodi di difficoltà produttiva, non essendo nelle condizioni di sostenere il costo del lavoro e di svolgere in modo remunerativo attività produttiva a pieno regime. Attualmente vi sono tre istituti o forme che disciplinano l'erogazione di questo trattamento: la cassa integrazione ordinaria, quella straordinaria e i contratti di solidarietà.

La cassa integrazione guadagni ordinaria (Cigo), riferita all'industria e all'edilizia, è un trasferimento in denaro erogato ai lavoratori per mantenere il loro livello salariale in caso di riduzione o sospensione dell'attività lavorativa, indipendente dalla volontà dai lavoratori e dall'impresa, a seguito di un evento temporaneo (ad esempio difficoltà nel reperire materie prime, guasti ai macchinari, intemperie stagionali, ecc.). I lavoratori a cui è concessa sono dipendenti assunti, compresi gli apprendisti con regolare contratto, non è invece dovuta ai dirigenti e ai lavoratori a domicilio. La norma prevede che i lavoratori tutelati debbano possedere un'anzianità di almeno novanta giorni nella stessa impresa che ne fa richiesta. La cassa integrazione ordinaria può avere una durata massima di tredici settimane consecutive, prorogabili ogni tre mesi sino a un massimo di cinquantadue settimane nell'anno. Nel periodo di cassa integrazione, al lavoratore è corrisposto un trattamento pari all'80% rispetto alla sua retribuzione globale, cioè al compenso che sarebbe spettato al lavoratore qualora non fosse intervenuto nessun tipo di evento di cassa integrazione. L'integrazione salariale non può comunque superare gli importi massimali stabiliti per legge in un dato anno. La Cigo è autorizzata con la modalità del conguaglio dei contributi sociali che il datore di lavoro deve versare. Le somme anticipate in busta paga ai propri dipendenti vengono recuperate dall'impresa mediante denuncia mensile all'INPS e tracciate nell'archivio Uniemens tenuto dallo stesso Ente.

La cassa integrazione straordinaria (Cigs)[4] è concessa dal Ministero del Lavoro e delle Politiche Sociali ed è erogata direttamente dall'INPS. Tale trattamento è un ammortizzatore sociale utilizzato per aiutare le aziende che si trovano in gravi difficoltà produttive a sostene-

4 Per i riferimenti normativi si rimanda a: D.lgs. n.788/1945; L. 1115/1968; L. 164/1975, artt.1-2; L. 223/1991; L. 236/1993.

re i costi di una sospensione o una ridotta capacità lavorativa che si protrae nel tempo. Tra i motivi principali che richiedono l'intervento straordinario di cassa integrazione, vi sono: a) la crisi aziendale, b) la riorganizzazione aziendale e c) l'applicazione di contratti di solidarietà (Sol difensivo). Nel primo caso, il periodo di cassa integrazione può durare non più di dodici mesi, consecutivi o meno; nel secondo il periodo massimo d'integrazione è pari a ventiquattro mesi (anche non consecutivi); infine, nel terzo caso può avere una durata di non più di ventiquattro mesi consecutivi. L'integrazione salariale, è calcolata con la stessa percentuale dell'ordinaria (80% della retribuzione globale) e non può superare gli stessi importi massimali. Anche per essa è previsto il recupero delle somme anticipate dall'impresa mediante denuncia mensile all'INPS, anche queste tracciate nell'archivio Uniemens.

Per ciò che riguarda i contratti di solidarietà (Sol), questi rappresentano un ammortizzatore sociale simile alla cassa integrazione ma che trovano un'apposita disciplina (L. 863/1984 artt. 1 e 2). Essi si suddividono in due tipologie: difensivo o rientranti (tipo A) ed espansivo o non rientranti (tipo B). I primi sono utilizzati, nel caso in cui vi sia una crisi aziendale e l'azienda stessa si trova nella necessità di ridurre il personale e quindi, anziché ricorrere ai licenziamenti, fa ricorso all'uso dei contratti di solidarietà, che comportano una perdita di retribuzione dovuta alla riduzione dell'orario di lavoro per tutti i lavoratori, attuando così una distribuzione degli svantaggi della crisi in egual misura su tutti. I secondi invece, si hanno quando l'azienda desidera fare nuove assunzioni, e decide di applicare una riduzione all'orario di lavoro al personale impiegato già nel processo produttivo, applicandogli una riduzione nelle retribuzioni. Il decreto che attesta la concessione dei contratti di solidarietà è previsto entro trenta giorni dalla richiesta dell'azienda. L'integrazione salariale da parte Inps dei contratti di solidarietà difensivi è pari al 60% della retribuzione persa. Mentre il contributo dei contratti di solidarietà espansivi garantisce al lavoratore il 25% della retribuzione persa a causa della riduzione del suo orario di lavoro.

10.3 Ricognizione delle fonti dati e strategie di rilevazione della cassa integrazione anticipata

Al fine di comprendere appieno l'approccio attraverso cui si stima la cassa integrazione anticipata dai datori di lavoro è indispensabile una breve disamina dei metadati della fonte utilizzata: l'archivio Uniemens dell'INPS. Dal punto di vista concettuale ciascuna denuncia aziendale è riorganizzata in diversi moduli (sezioni) con contenuti informativi diversi quali: le caratteristiche demografiche e professionali (nome, cognome, matricola, codice fiscale, tipologia di qualifica, tipo di contribuzione) di ciascun lavoratore e le posizioni retributive e contributive ad esso associate (INPS, 2015).

Dal punto di vista applicativo, invece, ciascuna fornitura dell'archivio Uniemens risulta composta di 12 tabelle delle quali, per gli scopi del presente lavoro, abbiamo preso in considerazione alcune informazioni presenti nelle seguenti tre:

1. EMENS: i) tipo di copertura della retribuzione settimanale (totale/parziale/nulla); ii) tipologia di evento coperto; iii) imponibile previdenziale mensile; iv) retribuzione mensile teorica (in assenza di eventi CIG, malattia, maternità, ecc.);
2. EMENS_CONGUAGLI: i) importi messi a conguaglio per le differenti tipologie di cassa integrazione; ii) numero di ore a copertura della CIG;
3. EMENS_DIFFACCRE: differenza/accredito, dichiarato dall'azienda come valore retributivo da considerare, nelle settimane in cui è avvenuto l'evento, a titolo di contribuzione figurativa.

Le tipologie di evento considerate nella presente ricerca comprendono la cassa integrazione guadagni ordinaria (Cigo), straordinaria (Cigs) e i contratti di solidarietà (Sol)[5]. Operativamente si è proceduto a selezionare gli individui campione dell'indagine IT-SILC (edizione 2016) all'interno del collettivo dei lavoratori presenti negli archivi INPS e, successivamente, ad applicare tecniche di data-fusion tra i due database selezionati.

Prima di illustrare le scelte procedurali dell'innovazione introdotta per la determinazione degli importi legati ai suddetti eventi, si intende spiegare la motivazione per cui approcci alternativi, seppure oggetto di sperimentazione, siano stati abbandonati in favore di quello messo a regime.

In una iniziale sperimentazione si è tentato di utilizzare la tabella EMENS_CONGUAGLI, dalla quale sono state ricavate tutte le informazioni indicanti gli importi messi a conguaglio per le differenti tipologie di cassa integrazione, nonché il numero di ore interessate da questo evento. Dalla tabella EMENS, invece, sono state estratte le indicazioni su: i) tipo di copertura settimanale (totalmente/parzialmente retribuita o non retribuita); ii) tipologia di evento coperto per tutte le settimane da calendario; iii) imponibile previdenziale mensile; iv) retribuzione mensile teorica, cioè il salario che il lavoratore avrebbe percepito in un determinato mese se non fosse intervenuto alcun tipo di evento (Cig, malattia, maternità ecc.). In questa stessa base dati sono stati anche considerati i codici identificativi della settimana e tutte le variabili riferite ai giorni lavorati o meno nel mese. Questo approccio, seppure promettente per il riferimento diretto alle somme anticipate dalle aziende, si è dimostrato impraticabile ai fini della costruzione delle componenti di reddito. L'evidenza empirica ha mostrato infatti che l'importo conguagliato supera spesso anche di molto la retribuzione teorica mensile riferita allo stesso mese di calendario, denotando un forte disallineamento temporale tra il momento dell'anticipo della prestazione e quello della richiesta di conguaglio (riferito anche a più periodi) da parte dello stesso datore di lavoro.

Una seconda sperimentazione per la stima della cassa integrazione è stata di tipo indiretto e si è avvalsa della ricostruzione dei periodi totalmente o parzialmente non retribuiti a causa dell'evento esaminato (Cigo, Cigs, Sol) a cui è associata una riduzione di retribuzione e quindi un trattamento a copertura. In primo luogo si è proceduto alla selezione di tutte le settimane da calendario "totalmente non retribuite" legate a un singolo evento tra quelli esaminati. Successivamente, per la ricostruzione del periodo associato agli eventi con "parziale copertura" settimanale[6] si è proceduto a utilizzare l'informazione sulla presenza o meno del lavoratore sul luogo di lavoro per ogni giorno del calendario. Tale abbinamento ha permesso di conteggiare esattamente per ciascun evento quanti giorni della settimana sono totalmente retribuiti e quanti totalmente non retribuiti e quindi di discriminare la casistica delle settimane totalmente non retribuite da quelle con parziale retribuzione. In quest'ultimo insieme sono state infine inserite anche le settimane in cui vi è una parziale copertura su tutti i giorni, ricavate per differenza tra l'insieme delle settimane interessate da uno degli eventi e quelle già definite a totale o parziale retribuzione. Replicando tale operazione per tutti i mesi dell'anno si giunge ad una stima del numero di settimane totalmente o parzialmente non retribuite nell'anno per evento considerato. Il limite di questo approccio consiste

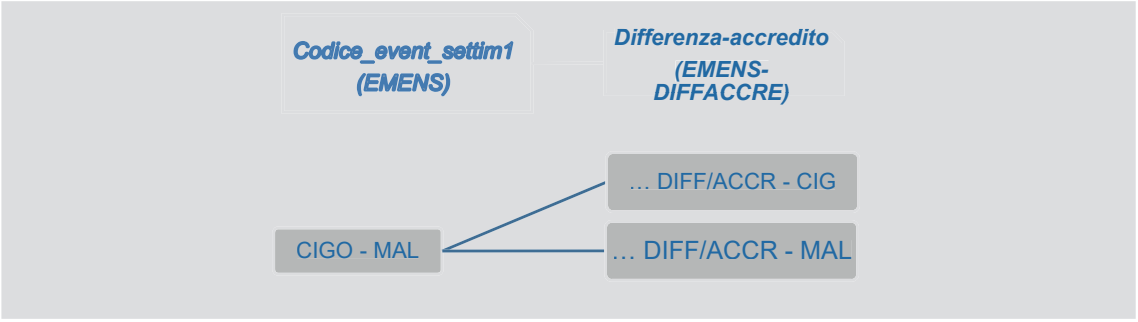
5 La cassa integrazione in deroga, pur essendo presente nell'archivio INPS, non figura nell'elenco dei trattamenti esaminati poiché per essa è previsto un complesso quadro normativo e sarà quindi oggetto di un futuro studio di fattibilità.

6 Una parziale copertura si verifica quando alcuni giorni della settimana sono retribuiti e altri completamente non retribuiti, oppure quando tutti giorni della settimana sono parzialmente retribuiti a causa di una riduzione giornaliera dell'orario di lavoro.

nel fatto che all'interno della stessa settimana possono presentarsi contemporaneamente più tipologie di eventi che si aggiungono alla cassa integrazione e ai contratti di solidarietà (malattia, maternità, congedi straordinari, ecc.). In tali circostanze risulta impossibile separare eventi estranei all'integrazione salariale nel conteggio delle settimane a parziale e totale copertura.

Una terza sperimentazione che è stata valutata come la soluzione più idonea al problema della stima della prestazione in esame, consiste nell'impiego della variabile "differenza/accredito" quale *proxy* dell'importo dell'integrazione salariale anticipata dal datore. Tale variabile, contenuta nell'apposita tabella denominata EMENS_DIFFACCRE, denota il valore della "retribuzione persa" dal lavoratore in conseguenza dell'evento (cassa integrazione, contratto di solidarietà, ecc.). In termini equivalenti, essa rappresenta l'ammontare della retribuzione che sarebbe stata considerata ai fini della contribuzione se il lavoratore avesse svolto la sua attività lavorativa durante il periodo in cui è stato assente per l'evento. In questi casi è l'azienda a dichiarare la variabile "differenza/accredito", cioè il valore retributivo da registrare nelle settimane in cui è avvenuto l'evento a titolo di contribuzione figurativa. Tale differenza rapportata alla retribuzione mensile teorica offre, come vedremo in seguito, l'elemento di base per il calcolo della prestazione oggetto di studio. Inoltre essa si riferisce sia ai casi di cig parziale che a quelli a zero ore. Affinché questa procedura possa essere applicata con successo, è necessario che ad ogni singolo evento presente nella tabella EMENS corrisponda un valore distinto di differenza/accredito della tabella EMENS_DIFFACCRE relativo allo stesso evento. In presenza di più eventi distinti nella stessa settimana, la retribuzione effettiva del lavoratore sarà costituita in parte dalla retribuzione pagata nel periodo lavorato e in parte dalla differenza/accredito risultante dagli eventi sopravvenuti che rappresenta la retribuzione persa (figura 10.1).

Figura 10.1 - Corrispondenza tra elementi delle due tabelle: eventi da calendario e differenza/accredito

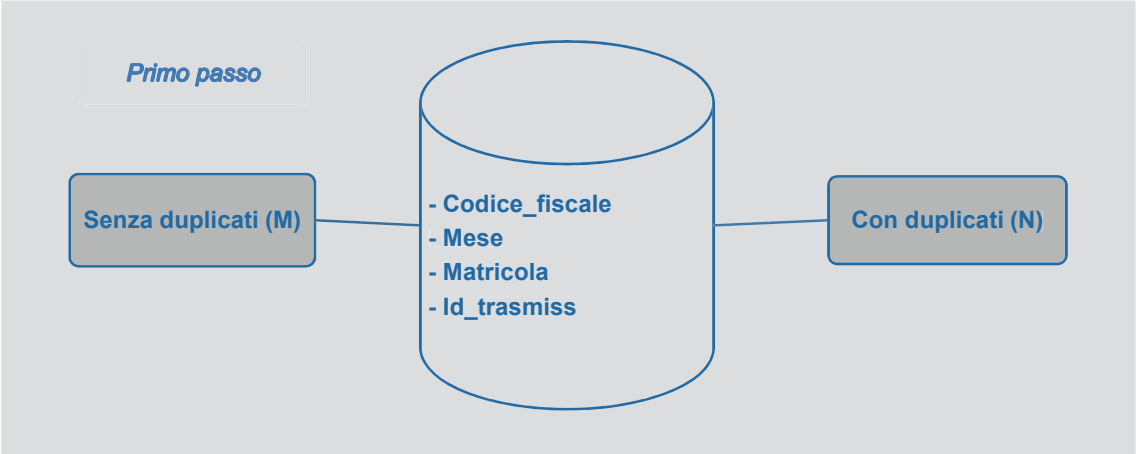


10.4 Trattamento preliminare dei dati di archivi amministrativi per l'uso della variabile differenza/accredito

Prima di procedere al calcolo della cassa integrazione è necessario identificare ed eliminare i potenziali errori insiti nell'uso degli archivi (EMENS, EMENS_DIFFACCRE). Solo successivamente si realizza un'integrazione tra le informazioni delle diverse fonti attraverso tecniche di *exact matching*. In questo paragrafo verranno illustrate alcune operazioni preliminari di *data cleaning*. Una delle casistiche più frequenti, quando si utilizzano dati amministrativi, è la presenza di duplicazioni da identificare sulla base di una chiave primaria. In questa applicazione sono state selezionate le posizioni lavorative dichiarate in un determi-

nato punto temporale (mese) mediante la chiave multipla: <codice_fiscale-id_trasmissione-mese-matricola>. Sulla base di questa chiave si sono generati due sottoinsiemi: i) il primo senza duplicati ii) il secondo con duplicati.

Figura 10.2 - Identificazione dei duplicati sulla base della chiave multipla iniziale da EMENS



Per quanto riguarda il primo insieme, l'assenza di duplicati implica che esso si caratterizza per casi che presentano per ciascun dipendente e all'interno dello stesso mese un'unica "id trasmission" di conseguenza tale variabile può essere eliminata in quanto il suo contenuto informativo è ridondante.

Se alla struttura identificativa del primo sottoinsieme **M** (senza duplicati) si omette il codice di trasmissione, si ottiene una nuova partizione dove sono a loro volta presenti sottoinsiemi che fanno registrare casi con (**CP**) e senza (**SP**) duplicazioni sulla base della seguente combinazione di codici: <codice_fiscale-mese-matricola>. Per quanto riguarda il primo sottoinsieme (**CP**) le duplicazioni sono da attribuire alla presenza di più eventi per ciascun dipendente e mese che generano più di un importo mensile della retribuzione teorica. La strategia adottata è stata quella di procedere ad identificare per ciascun singolo evento (Cigo, Cigs, Sol) la frequenza con cui si presentano i diversi importi mensili della retribuzione teorica in capo allo stesso soggetto. In caso di frequenza multipla di uno stesso importo si procede a selezionare il valore massimo (come rappresentativo) collassando la struttura in un'unica riga. In caso di frequenza unitaria si sommano gli importi distinti della retribuzione teorica (Freq=1) che si presentano lungo una stessa struttura <codice_fiscale-mese-matricola>. Ciò che si ottiene è un sottoinsieme senza importi duplicati rispetto alla chiave <codice_fiscale-mese-matricola>.

Figura 10.3 - Identificazione e trattamento duplicati su chiave ridotta (omettendo identificativo trasmissione)

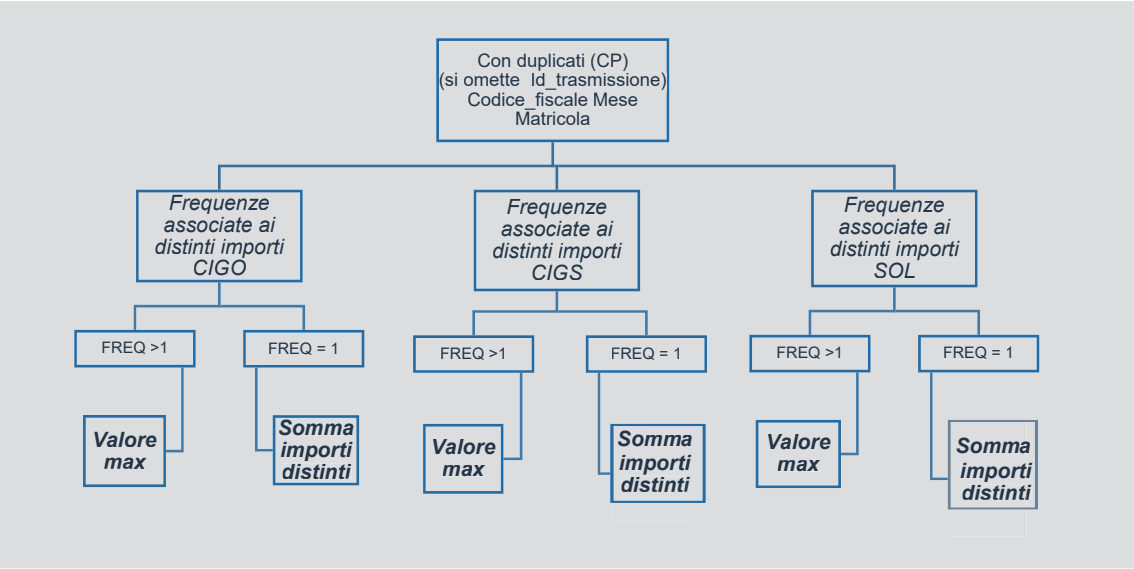


Figura 10.4 - Esempio di trattamento di importi multipli in caso di duplicazione della chiave ridotta

Codice_fiscale	Mese	Matricola	Event_set1	…	Event_set6	Retrib_teor
AAA…	LUG	XXX	CIGS	…	CIGS	RT1
AAA…	LUG	XXX	CIGO	…	CIGO	RT2
AAA…	LUG	XXX		…	CIGO	RT3
AAA…	LUG	XXX				RT1+ RT2+RT3

Nel caso del sottoinsieme **SP** (senza duplicazioni), invece, la retribuzione teorica presa in considerazione è l'unica presente per ogni chiave.

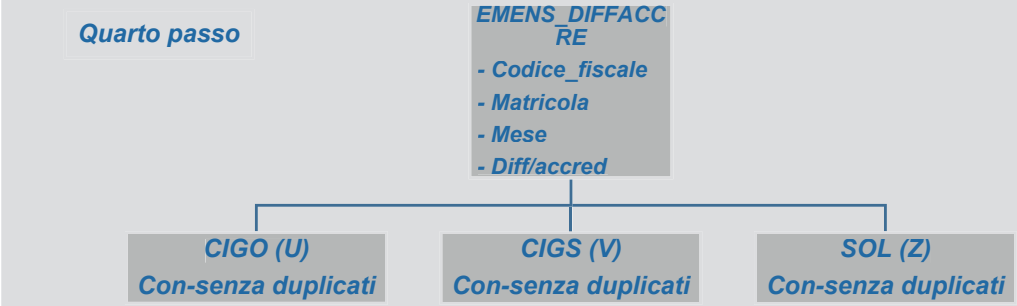
In riferimento al sottoinsieme **N**, di cui alla Figura 10.2, i duplicati sono dovuti alla circostanza che nello stesso periodo si possono associare allo stesso individuo più tipologie di eventi a cui si corrispondono importi di retribuzione mensile teorica talora diversi o, in altri casi, coincidenti. Per ogni casistica, si estrae il sottoinsieme di soli eventi di cassa integrazione (Cigo e Cigs) o di contratti di solidarietà (Sol) da associare ad un unico importo di retribuzione mensile teorica (Terzo passo). Una volta filtrati i soli eventi di interesse (Cigo/Cigs e Sol) in corrispondenza del sottoinsieme **N** si hanno solo record unici in base alla chiave: <codice_fiscale-id_trasmissione-matricola-mese>; pertanto si procede a selezionare l'unico importo corrispondente di retribuzione teorica mensile. Nel caso vi siano più importi distinti si procede alla somma (figura 10.5).

Figura 10.5 - Esempio di duplicati presenti nella chiave multipla iniziale (sottoinsieme N)

Cod_fisc	Mese	Matr	Id_trasmis	Event_set1	.....	Event_set6	Retrib_teor.	Imponib_previd.
BBB...	DIC	YYY	1237	SOL			RT4	IP4
BBB...	DIC	YYY	1237		SOL	SOL	RT5	IP5
BBB...	DIC	YYY	1237	SOL			RT4	IP4

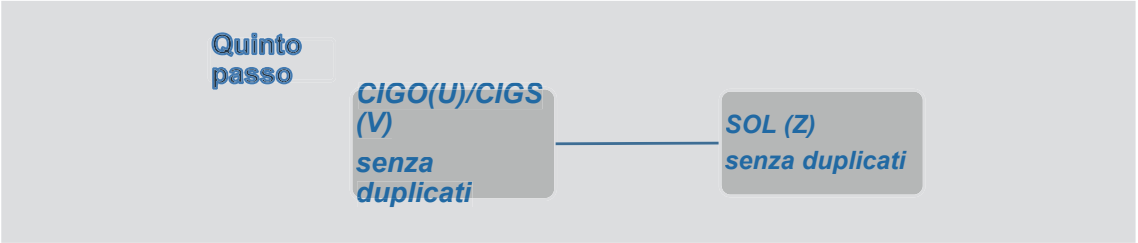
Quest'ultima operazione, conclude le fasi di *data cleaning* con l'identificazione di valori unici per periodo considerato nella tabella EMENS. A partire dalla tabella EMENS_DIFFACCRE, si procede ad estrapolare il sottoinsieme dei soli eventi distinti di cassa integrazione ordinaria (Cigo) e straordinaria (Cigs) e dei soli contratti di solidarietà (Sol), dove è presente la variabile differenza accredito con e senza duplicati, mediante la chiave multipla: <codice_fiscale-matricola-mese>.

Figura 10.6 - Trattamento dei duplicati in base alla chiave multipla da EMENS_DIFFACCRE



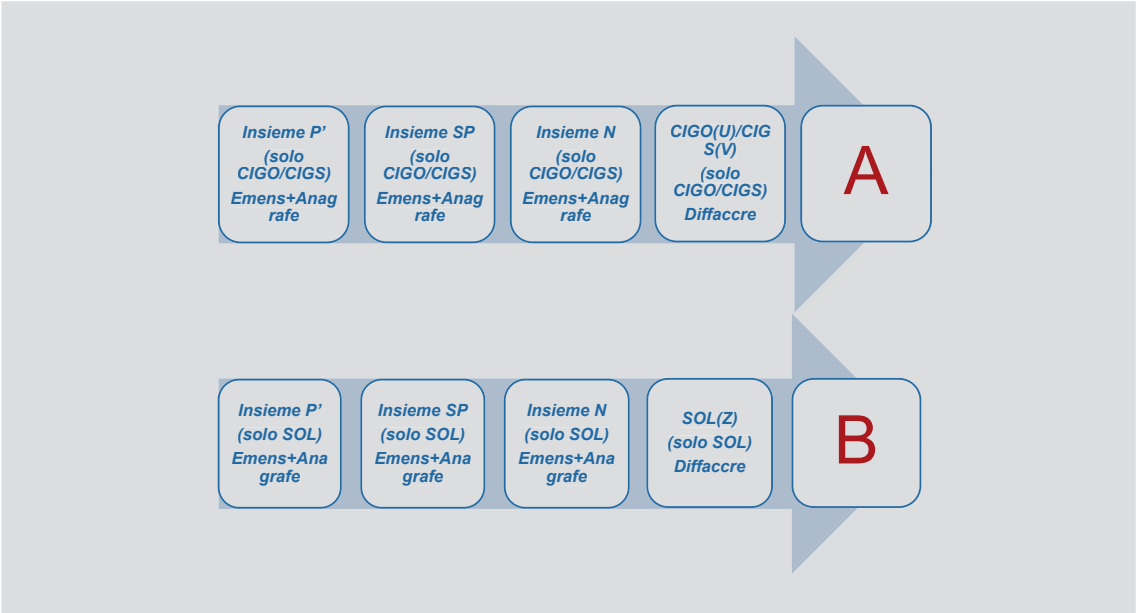
Nei casi di duplicazione degli eventi considerati, si procede ad ottenere un unico importo della variabile differenza/accredito.

Figura 10.7 - Creazione dei sottoinsiemi senza duplicati distinti per eventi: Cigo/Cigs e Sol



A questo punto è possibile unire tutte le tabelle (trattati con operazione di data cleaning), ottenute separatamente per i casi di cassa integrazione (Cigo, Cigs) e per quelli di solidarietà (Sol).

Figura 10.8 - Creazione dei sottoinsiemi senza duplicati distinti per eventi: Cigo/Cigs e Sol

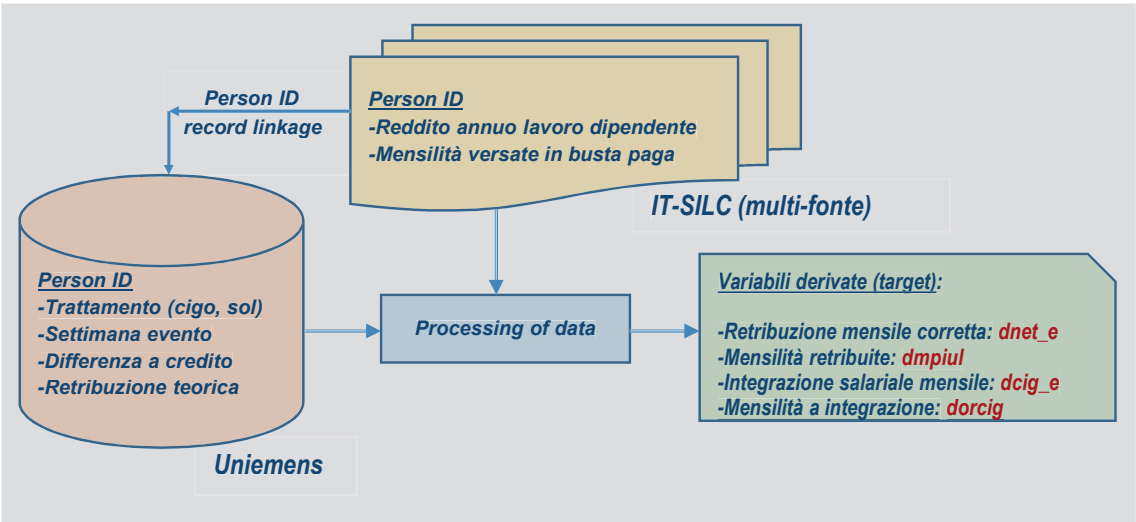


Le due tabelle **A** e **B**, risultanti dall'integrazione delle tue tabelle dell'archivio INPS, costituiscono la base sui cui applicare la formula di calcolo per la stima della cassa integrazione che tratteremo nel prossimo paragrafo.

10.5. Metodo di calcolo

Al fine di raggiungere l'obiettivo di piena copertura dell'integrazione salariale, includendo anche la componente anticipata dal datore di lavoro, è necessario formalizzare uno schema che permetta di fondere le informazioni elementari di natura previdenziale (Uniemens) con quelle elaborate in IT-SILC. La seguente rappresentazione grafica illustra la strategia adottata.

Figura 10.9 - Integrazione della base dati Uniemens ed IT-SILC per la stima della CIG anticipata



Il primo passo verso l'integrazione si realizza abbinando le unità delle due fonti EU-SILC (IT-SILC) e Uniemens, attraverso la chiave di aggancio (*matching key*) costituita dal codice fiscale della persona fisica (Person ID). Il successivo prevede l'integrazione vera e propria del contenuto informativo delle rispettive basi di dati (*data processing*) e la generazione delle variabili di interesse (variabili derivate), in analogia con il metodo descritto in Wallgren and Wallgren, 2017.

L'estrazione e l'elaborazione dei dati elementari di fonte previdenziale (Uniemens) si articola lungo le seguenti fasi: 1) selezione delle tipologie di trattamento di integrazione salariale (Cigo, Cigs, Sol) e settimane per le quali il lavoratore ne ha beneficiato; 2) ricostruzione dei mesi di fruizione della cassa integrazione a partire dalle settimane di calendario in cui accade l'evento; 3) calcolo della quota (media) di retribuzione persa a causa dell'evento cassa integrazione nel periodo di fruizione[7]. Dalla fonte IT-SILC si recuperano le informazioni di base relative al reddito netto annuo versato in busta paga al lavoratore dipendente e le corrispettive mensilità pagate. I dati di fonte previdenziale (Uniemens) sono dunque strumentali alla ripartizione del reddito in busta paga rilevato in IT-SILC nelle due componenti: retribuzione e integrazione salariale (prestazione sociale). La scomposizione tra le suddette componenti può essere schematizzata come segue:

$$\begin{aligned} \textit{reddito annuo busta paga} = {} & \textit{retribuzione annua teorica nel periodo senza cassintegrazione} + \\ & \textit{retribuzione annua nel periodo di riduzione oraria a cassintegrazione} + \\ & \textit{integrazione salariale annua} \end{aligned}$$

Formalizzando il discorso, l'espressione può essere riscritta come:

$$\begin{aligned} & Y_{dfs}^{A} = Y_{dt}^{M} \cdot (dmpiul - dmcig) + Y_{dt}^{M} \cdot (1 - dorcig) \cdot dmcig + Y_{dt}^{M} \cdot ccig \cdot dorcig \cdot dmcig = \\ & = Y_{dt}^{M} \cdot (dmpiul - dorcig \cdot dmcig) + Y_{dt}^{M} \cdot ccig \cdot dorcig \cdot dmcig = \\ & = Y_{dt}^{M} \cdot (dmpiul - dorcig \cdot (1 - ccig) \cdot dmcig) \end{aligned} \qquad [1]$$

Da cui si ottiene:

$$Y_{dt}^{M} = Y_{dfs}^{A} / (dmpiul - (1 - ccig) \cdot dorcig \cdot dmcig) \qquad [2]$$

Dove:

Y_{dfs}^{A} : reddito annuo versato in busta paga al lavoratore dipendente da fonte fiscale (IT-SILC);

Y_{dt}^{M} : retribuzione mensile da lavoro dipendente teorica in assenza dell'evento cig (variabile derivata);

dmpiul : numero di mesi retribuiti (IT-SILC);

dmcig : numero di mesi soggetti a cassa integrazione (Uniemens);

dorcig : quota parte della componente retributiva persa a causa della cassa integrazione (Uniemens); il suo complemento a 1 rappresenta invece la quota parte della busta paga associata alle sole ore lavorate, ovvero retribuita;

ccig : coefficiente di integrazione del reddito riconosciuta dall'INPS per trattamento (Uniemens).

A titolo di esempio, con un reddito annuo versato in busta paga al lavoratore (Y_{dfs}^{A}) di 13.000

7 Intesa come rapporto tra l'ammontare di retribuzione persa a causa dell'evento cassintegrazione (cioè la sommatoria della "differenza/ accredito" nei mesi in cui è valorizzata) e la retribuzione teorica spettante nello stesso periodo in assenza di qualunque evento.

euro, un numero di mesi pagati ($dmpiul$) uguale a 12, un periodo di cassa integrazione ($dmcig$) di 2 mesi, una percentuale di riduzione della componente retributiva a causa dell'evento *CIG* ($dorcig$) del 60%, e infine un coefficiente di integrazione salariale INPS ($ccig$) pari all'80%; si perviene, tramite l'espressione [2], a una retribuzione mensile teorica in assenza dell'evento *CIG* uguale a: $Y_{dt}^{M} = 13000/(12 - 0{,}2 \cdot 0.6 \cdot 2) = 1.105{,}4$ euro.

Una volta determinato il valore della retribuzione teorica mensile, le tre componenti reddituali spettanti al lavoratore si ottengono facilmente tramite l'equazione [1]:

$$retribuzione\ annua\ nei\ mesi\ senza\ cassintegrazione = Y_{dt}^{M} \cdot (dmpiul - dmcig) = 11.054{,}4\ €$$

$$retribuzione\ annua\ nel\ periodo\ di\ riduzione\ oraria\ per\ cig = Y_{dt}^{M} \cdot (1 - dorcig) \cdot dmcig = 884{,}4\ €$$

$$integrazione\ salariale\ annua = Y_{dt}^{M} \cdot ccig \cdot dorcig \cdot dmcig = 1061{,}2\ €$$

Il valore (medio) dell'integrazione salariale mensile ($dcig\_e$) che, unitamente alle mensilità in cui si presenta l'evento ($dmcig$), rappresenta la nostra variabile obiettivo, si ottiene semplicemente dal rapporto tra l'ammontare annuo dell'integrazione salariale anticipata e il periodo a cui si è fatto ricorso (numero mensilità). Nell'esempio citato l'integrazione salariale (media) mensile è pari a 530,60 euro. Il valore della retribuzione media mensile ($dnet\_e$) si ricava sommando le due componenti retributive, rispettivamente nei periodi con o senza cassa integrazione, e dividendo il totale per i mesi retribuiti ($dmpiul$). Nel esempio di cui sopra essa corrisponde a 994,90 euro.

Il procedimento di scomposizione risulta lievemente più complesso qualora si intenda determinare il valore della tredicesima o relativo rateo. La normativa vigente stabilisce che, nel caso di CIG a zero ore (assenza totale di prestazione lavorativa nel periodo), il pagamento dell'integrazione salariale grava sull'INPS nella misura dell'80% della retribuzione globale che sarebbe spettata allo stesso per le ore di assenza di CIG, maggiorate del rateo intero di tredicesima (nel rispetto dei massimali mensili previsti). Nel caso poi di CIG con riduzione parziale di orario lavorativo, l'INPS provvede ad erogare l'integrazione salariale inclusiva dei ratei di tredicesima per la sola parte di assenza di prestazione lavorativa, mentre al datore di lavoro spetta, oltre alla retribuzione, la quota di tredicesima a suo carico per le ore lavorate[8].

La formula di calcolo dell'integrazione salariale resta identica al caso precedente, poiché essa implicitamente include i ratei di tredicesima, mentre la suddivisione della retribuzione annua globale (nei periodi con o senza *CIG*) tra le componenti "mensile corrente" e "tredicesima" deve essere riparametrata come di seguito:

$Y_{DCIG}^{A} = Y_{dt}^{M} \cdot ccig \cdot dorcig \cdot dmcig = 1061{,}2\ euro$ (integrazione salariale annua);

$DCIG\_E = Y_{dt}^{M} \cdot cig \cdot dorcig = 530{,}6\ euro$ (integrazione salariale mensile);

$Y_{DIP\_NT}^{A} = \left(Y_{dfs}^{A} - Y_{DCIG}^{A}\right) \cdot \left(\frac{12}{13}\right) = 11.020{,}4\ euro$ (retribuzione annua al netto della tredicesima);

$DNET\_E = \frac{Y_{DIP\_NT}^{A}}{dmpiul} = 918{,}4\ euro$ (retribuzione mensile corrente al netto della tredicesima o rateo);

8 Per semplicità di calcolo non si tiene conto dell'ulteriore specifica prevista dalla normativa, secondo cui il datore di lavoro è tenuto a integrare quella parte di ratei di tredicesima fino a concorrenza dei massimali INPS, ove questi ultimi non siano superati.

$DALCOA\_E = \left(Y^A_{dfs} - Y^A_{DCIG}\right) \cdot \left(\frac{1}{13}\right) = 918{,}4\ euro$ (tredicesima o relativo rateo sulla quota retributiva).

Si fa notare che la procedura di calcolo dell'integrazione salariale mensile deve essere tale da non generare valori inconsistenti rispetto ai massimali erogabili secondo legge (CIG_{max}). Per scongiurare tale eventualità si impone il seguente vincolo:

$$se\ DCIG\_E > CIG_{MAX} \quad => \quad \begin{cases} \check{Y}^A_{DCIG} = CIG_{MAX} \cdot dmcig \\ \check{Y}^A_{DIP\_NT} = \left(Y^A_{dfs} - \check{Y}^A_{DCIG}\right) \cdot \left(\frac{12}{13}\right) \\ DALCOA\_E = \left(Y^A_{dfs} - \check{Y}^A_{DCIG}\right) \cdot \left(\frac{1}{13}\right) \end{cases}$$

10.6. Risultati dell'implementazione

L'implementazione della nuova procedura per il calcolo della cassa integrazione, fondata sui dati amministrativi applicati all'indagine IT-SILC 2016, ha consentito di rilevare ben 574 lavoratori beneficiari di integrazione salariale anticipata dal datore di lavoro per conto dell'INPS. Per essi è stato possibile scorporare il valore della prestazione sociale dalla corrispettiva retribuzione. Rispetto ai risultati dell'edizione 2015, dove tramite intervista diretta figuravano appena 93 titolari, la nuova strategia di rilevazione fornisce un netto guadagno informativo. Applicando i coefficienti di riporto all'universo per l'indagine IT-SILC 2016 si stima che nel 2015 vi siano 700 mila lavoratori beneficiari di cassa integrazione anticipata, per un importo globale di 1,3 miliardi di euro. Purtroppo, non avendo a disposizione una fonte esterna (benchmark) che consente la validazione "diretta" delle stime sul numero dei titolari e sulla relativa spesa anticipata, si può ottenere solo una validazione indiretta della spesa erogata tramite altre fonti. In base al conto satellite della protezione sociale (macro dato) si osserva come nell'anno 2015 l'importo complessivamente versato per le integrazioni salariali ammonti a 2,5 miliardi di euro, di cui 700 milioni in pagamenti diretti dell'INPS (archivio micro dati trattamenti non pensionistici). Per differenza si ottiene dunque un ammontare di spesa anticipata per integrazioni salariali stimata in 1,8 miliardi di euro. La nuova metodologia, pur fornendo una stima di spesa non perfettamente allineata al valore del benchmark (stima indiretta), segna tuttavia un notevole passo in avanti in termini di qualità del dato. Ciò soprattutto se confrontato con il risultato dell'edizione precedente anno 2014 (sola intervista diretta), dove l'importo stimato per integrazioni anticipate era pari a 127 milioni di euro contro il valore di benchmark (fonti micro-macro) pari a 2,2 miliardi di euro. Per il futuro si prevede di migliorare e affinare la tecnica utilizzata, con l'ulteriore inclusione della componente relativa all'integrazione salariale in deroga, tralasciata per ragioni di complessità nella riproduzione del quadro normativo vigente.

11. UN'ANALISI ESPLORATIVA DELL'ARCHIVIO INPS SU CERTIFICAZIONI TELEMATICHE DI MALATTIA[1]

11.1 Introduzione

Il presente lavoro mira a verificare le potenzialità informative dell'archivio INPS sulle certificazioni telematiche di malattia ai fini della ricostruzione della variabile target EU-SILC: "*Sickness benefits (PY120G/N)*", attualmente non rilevata nell'indagine italiana sui redditi. Infatti, vi è da osservare come questa componente monetaria, costituita in prevalenza dalle indennità di malattia erogate ai lavoratori dipendenti, sia oggettivamente difficile da rilevare tramite intervista diretta, anche di fronte all'intervistatore più esperto e al rispondente più istruito. Volendo pure formulare una domanda diretta sull'importo annuo percepito per l'indennità di malattia, l'intervistato non è in grado di riportare un valore certo, poiché tale voce economica non è facilmente individuabile all'interno del cedolino dello stipendio. Del pari complicata è la ricostruzione della componente monetaria tramite il quesito indiretto, ovvero chiedendo il numero di giorni di malattia presi nell'anno di riferimento. Risulta evidente la difficoltà di recuperare un'informazione qualitativamente valida quando lo scarto temporale, tra il momento in viene somministrato il quesito all'intervistato e l'insorgenza del fenomeno (malattia), supera addirittura l'anno. In letteratura sono note le distorsioni legate all'effetto memoria e all'effetto *telescoping* che inficiano la qualità del dato. Nel primo caso, quando il periodo di riferimento è molto ampio, gli eventi vengono dimenticati; mentre nel secondo caso può accadere che gli eventi avvenuti prima, ma ridosso dell'inizio del periodo di riferimento, vengano riportati comunque dai rispondenti, con un evidente effetto di sovrastima (Bagatta, 2006). In ragione di queste difficoltà, la Commissione europea ha concesso la possibilità agli Istituti Nazionali di Statistica di ricollocare (impropriamente) l'indennità di malattia tra le voci che compongono la retribuzione.

"Employee cash or near cash income includes … payments made by employers to an employee in lieu of wages and salaries through a social insurance scheme when unable to work through sickness, disability or maternity leave where such payment cannot be separately and clearly identified as social benefits" (Eurostat, 2017).

L'obiettivo ultimo della sperimentazione è l'individuazione di fonti, alternative alla raccolta diretta, per la stima delle prestazioni sociali che ricadono nella funzione/rischio malattia e, in subordine, per la corretta rilevazione della variabile retributiva. Questo lavoro rientra nel filone della ricerca sperimentale, sviluppata all'interno del team italiano IT-SILC, avente come scopo la progettazione e l'implementazione di nuovi e più efficienti i metodi

1 L'autore di questo capitolo è Paolo Consolini.
Si desidera ringraziare i referenti statistici del Coordinamento Generale Statistico Attuariale dell'INPS per la collaborazione offerta nell'ambito dei progetti di ricerca del Psn, finalizzata alla messa a disposizione dell'Archivio sui Certificati di malattia. Un ringraziamento particolare va alle colleghe Ilaria Girau e Laura Ghezzi del Servizio RDD dell'Istat che hanno organizzato gli incontri tecnici inter-istituzionali e fornito supporto amministrativo-gestionale.

di stima per le componenti di reddito, con l'ausilio di fonti informative alternative. Il ricorso al dato amministrativo rappresenta, a giudizio dell'autore, la via principale per il conseguimento di validi risultati sperimentali e per l'arricchimento dell'informazione statistica. Tuttavia, prima di procedere a una qualsiasi misurazione di un fenomeno sociale occorre: a) definire con esattezza il fenomeno che si intende osservare; b) esaminare le fonti che contengono l'informazione di interesse; c) individuare i metodi appropriati di calcolo (o di stima); e infine d) verificare empiricamente la bontà dei vari metodi a confronto. Il presente documento è articolato in sei paragrafi, nei successivi quattro verranno nell'ordine illustrati i punti cardine sopramenzionati, l'ultimo è dedicato alle conclusioni.

11.2 Malattia: identificazione dell'unità di analisi, aspetti concettuali e classificazioni

Le prestazioni sociali che ricadono sotto la funzione Malattia (classificazione ESSPROS)[2], comprendono sia prestazioni in denaro che compensano integralmente o in parte la perdita di reddito per inabilità al lavoro temporanea (dovuta a malattia o infortunio), sia prestazioni in natura di assistenza sanitaria per mantenere, ripristinare e migliorare la salute delle persone protette (Consolini, 2000). Il progetto europeo EU-SILC ha come finalità ultima quella di quantificare l'importo delle sole prestazioni in denaro che ricadono nella funzione malattia, come peraltro nelle restanti funzioni di protezione sociale. In particolare, nel contesto italiano (IT-SILC) quelle in denaro a tutela del rischio di malattia sono rappresentate da: 1) l'indennità giornaliera per malattia nel contesto della tutela economica della malattia generica; 2) l'indennità giornaliera (corrisposta durante il ricovero o la cura ambulatoriale), l'indennità giornaliera post-sanatoriale, l'assegno mensile di cura e l'assegno speciale di gratifica, tutte nell'ambito dell'assicurazione contro la tubercolosi; 3) l'indennità giornaliera per inabilità temporanea, quale specifica tutela economica di tipo non pensionistico contro gli infortuni sul lavoro.

L'unità di analisi della presente ricerca sperimentale è data dall'indennità di malattia, di cui al precedente punto 1, intesa come prestazione sociale individuale. Si definisce prestazione individuale il complesso dei trattamenti individuali erogati, per uno stesso titolo e funzione, al medesimo beneficiario da parte di uno o più centri di spesa in un determinato arco temporale[3]. Come vi evince da questa prima disamina, la trattazione dell'argomento relativo all'unità analisi è alquanto complessa e merita un graduale livello di approfondimento. Riprendendo uno schema concettuale dell'autore pubblicato in un precedente lavoro (Consolini e De Carli, 2002), l'indennità di malattia si compone di otto distinti trattamenti monetari non pensionistici che differiscono in base al centro erogatore di spesa e la modalità di erogazione (figura 11.1).

L'elenco dei trattamenti include: 1. erogazioni dirette dell'INPS (regime pubblico); 2. erogazioni dirette dell'ex-IPSEMA ora INAIL (regime pubblico); 3. erogazioni interamente a carico del datore di lavoro privato (regime privato); 4. erogazioni in regime privato a carico del datore di lavoro pubblico: amministrazione centrale; 5. erogazioni in regime privato a carico del datore di lavoro pubblico: amministrazione locale; 6. erogazioni in regime privato a carico del datore di lavoro pubblico: enti di previdenza; 7. anticipazioni del datore di lavoro per conto dell'INPS (regime pubblico); 8. Integrazioni alla prestazione di base previste dai CCNL, per semplicità indicate come retribuzioni ridotte (regime privato).

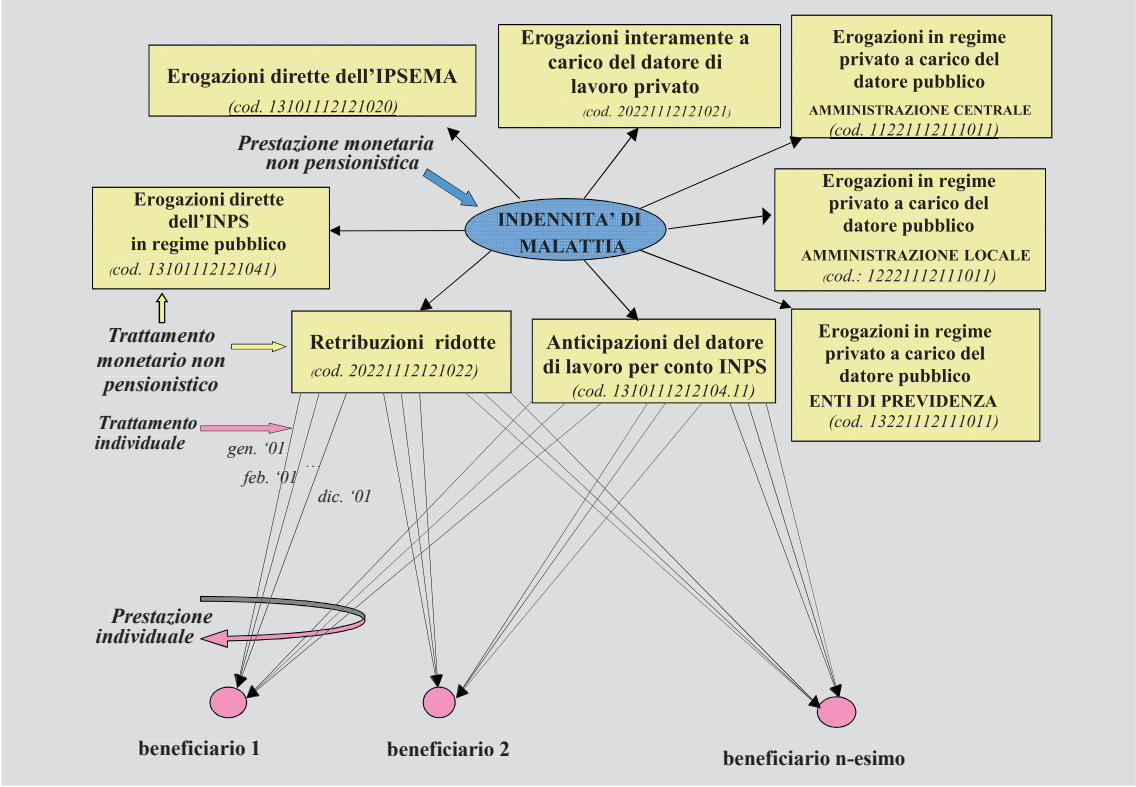
2 In base al rischio o bisogno prevalente coperto da ciascun regime di protezione sociale.

3 Per la definizione dei concetti relativi a prestazione/trattamento monetario non pensionistico si rimanda al contributo di Consolini e De Carli, 2002.

In merito al punto 1 la norma stabilisce che, in determinate circostanze, l'indennità di malattia sia di stretta competenza dell'INPS, con pagamento diretto al lavoratore avente diritto (regime pubblico). In particolare, l'erogazione diretta da parte dell'Istituto spetta a:

- lavoratori agricoli a tempo determinato (OTD);
- lavoratori stagionali;
- lavoratori il cui datore di lavoro non è in grado di anticipare l'indennità;
- lavoratori disoccupati o sospesi che non usufruiscono della Cassaintegrazione;
- dipendenti di aziende sottoposte a procedura concorsuale: fallimento, concordato preventivo, amministrazione controllata, liquidazione coatta amministrativa, amministrazione straordinaria;
- lavoratori che ricevono il pagamento diretto della Cigo, Cigs o CIG (in deroga);
- lavoratori aventi diritto all'erogazione diretta secondo disposizioni della DTL (Direzione Territoriale del Lavoro);
- dipendenti che si sono ammalati prima che l'azienda cessasse l'attività;
- dipendenti il cui il datore di lavoro si rifiuti di anticipare l'indennità (l'INPS è tenuta a diffidarlo e procedere al pagamento diretto qualora il datore non abbia ottemperato nei termini).

Figura 11.1 - Prestazione monetaria non pensionistica e componenti: indennità di malattia



In relazione punto 3, dove figurano le prestazioni interamente a carico del datore di lavoro privato, si osserva come l'indennità di malattia venga comunque erogata, in regime privato, a favore di determinate figure professionali[4]: impiegati, quadri e dirigenti dell'industria, del credito e dell'agricoltura. Restano, invece, totalmente esclusi dal beneficio i portieri e dipendenti

4 La materia è disciplinata dai Contratti Collettivi Nazionali del Lavoro (CCNL).

da proprietari di stabili, i viaggiatori e piazzisti, e i dipendenti da partiti politici e associazioni sindacali. Per tutti gli assicurati INPS (di cui al punto 1) la norma prevede, poi, che i primi tre giorni di malattia, la cosiddetta "carenza", siano a totale carico del datore di lavoro. Il periodo di carenza rientra a pieno titolo nella fattispecie appena discussa.

Nel caso dei dipendenti pubblici (punti 4-5-6), il pagamento dell'indennità di malattia spetta alla stessa amministrazione presso cui lavora il dipendente (regime privato). Non vi è modo di individuare la componente economica della malattia nella busta paga del lavoratore, per cui l'unica via per ottenere una stima è quella indiretta, cioè tramite le giornate di assenza per malattia e la retribuzione media giornaliera di base.

Per tutti i restanti assicurati INPS (punto 7), l'erogazione della prestazione avviene tramite il datore di lavoro che anticipa in busta paga il pagamento di competenza dell'INPS, conguagliandolo successivamente con i contributi sociali da versare allo stesso Ente previdenziale.

Per motivi di semplicità di trattazione e reperibilità delle fonti statistiche, il campo di osservazione è stato circoscritto ai primi sette trattamenti, con l'esclusione delle integrazioni alla prestazione di base fornite dal datore privato sulla base dei CCNL (punto 8). Il livello di copertura dei trattamenti qui esaminati, sia in termini di beneficiari che di importi ricevuti/erogati, è pressoché esaustivo.

11.3 Malattia: le fonti del dato statistico

Le fonti del dato statistico in esame si legano più strettamente al concetto di "trattamento" non pensionistico di quanto non accada al livello di "prestazione", essendo il primo una componente elementare della seconda. Nelle fasi di studio delle fonti del dato statistico e nella relativa analisi dei metadati si osserva frequentemente la presenza di una corrispondenza biunivoca tra fonte e trattamento. Nel caso dell'indennità di malattia, le fonti amministrative che consentono la rilevazione diretta (importo del trattamento) o indiretta (tramite il numero di giorni indennizzati) del fenomeno sono rappresentate da:

a. Certificazioni telematica della malattia dei dipendenti pubblici (rilevazione indiretta dei trattamenti di cui ai punti 4-5-6);
b. Certificazioni telematica della malattia dei dipendenti del settore privato (rilevazione indiretta dei trattamenti di cui ai punti 1-3-7);
c. Archivio INPS prestazioni dirette non pensionistiche (rilevazione diretta del trattamento di cui ai punti 1 e 2)
d. Archivio INPS Emens delle differenze di accredito dei contributi figurativi (rilevazione diretta del trattamento di cui al punto 7 a livello mensile);
e. Archivio INPS Emens dei conguagli degli importi dei trattamenti anticipati dai datori per conto dell'INPS (rilevazione diretta del trattamento di cui al punto 7 a livello mensile);
f. Casellario INPS delle posizioni previdenziali attive (rilevazione diretta del trattamento di cui al punto 7 a livello annuale).

Da questa lista delle fonti trova conferma il fatto che, nei confronti dei lavoratori pubblici (lato destro di Figura 11.1), sia possibile la sola stima indiretta dell'indennità di malattia. Infatti, questi ultimi ricevono il beneficio direttamente in busta paga dal proprio datore di lavoro (AA.PP.) in regime privato, cioè senza il ricorso a un soggetto terzo (Ente di previdenza) che si frappone tra datore e lavoratore nell'erogazione del flusso monetario. Da rilevare, inoltre, la presenza di un certo grado di sovrapposizione tra le fonti nella rilevazione dei trattamenti di malattia. Così, ad esempio, l'archivio INPS sui certificati di malattia

dei dipendenti del privato (punto b.) concorre alla misurazione del fenomeno per lo stesso collettivo presente negli archivi INPS delle prestazioni dirette (punto *c.*), delle differenze di accredito contributivo (punto d.), degli importi a conguaglio (punto *e.*) e delle posizioni previdenziali attive (punto f.). Ciò che contraddistingue il primo dai rimanenti quattro archivi è la variabile osservata; nel primo caso si rileva, per ciascuna unità, il "numero dei giorni di malattia nell'anno" (desunta dai distinti periodi di malattia); mentre nelle restanti fonti si registra l'importo annuo erogato o la retribuzione persa per lo stesso evento[5]. Le fonti di cui a punti *d.* e *f.* sono accomunate sia dallo stesso collettivo di riferimento: dipendenti del privato; sia dalla medesima variabile di osservazione: retribuzione persa a causa dell'evento morboso. La compresenza di fonti che misurano il fenomeno di interesse su un identico collettivo (seppure sotto diverse angolazioni: variabili) permette tanto di realizzare il processo di validazione tra i dati, quanto di migliorare la qualità e arricchire il contenuto dei dati che saranno oggetto di futura pubblicazione. Come si vedrà nei successivi paragrafi, l'articolazione delle norme che regolano l'indennità di malattia è tale da rendere molto complesso sia il processo di validazione che quello di ricostruzione dei relativi importi o della durata dell'evento.

Segue una rassegna delle principali caratteristiche delle fonti amministrative utilizzate e relativi metadati.

11.3.1 Certificazione telematica della malattia dei dipendenti pubblici

L'archivio in questione consente di acquisire informazioni tanto sul beneficiario dell'indennità, quanto sulle caratteristiche del trattamento e del datore pubblico (in qualità di centro erogatore di spesa). In merito al beneficiario si dispone, unitamente al codice fiscale (ID), delle informazioni relative a: sesso, data e luogo di nascita (per la verità desumibili dal codice fiscale), seguite dal comune e provincia di residenza. In ordine al trattamento vengono riportate: il periodo della malattia (in termini di data di inizio evento e fine prognosi), la presenza o meno di causa di infortunio, la tipologia del periodo (inizio, continuazione, ricaduta), la tipologia di visita di accertamento (ambulatoriale, domiciliare, pronto soccorso, ricovero). Infine, riguardo al datore di lavoro, vengono comunicati i dati del codice fiscale dell'amministrazione pubblica e la sua descrizione (titolo o nome).

11.3.2 Certificazione telematica della malattia dei dipendenti del privato

Anche per questo archivio si rilasciano informazioni che riguardano nell'ordine: il beneficiario, il trattamento e il datore (in questo caso privato). In merito al primo, oltre al codice fiscale (ID) figurano i dati della qualifica professionale e del settore di lavoro dell'assicurato, mentre si omette la sua residenza. In relazione al trattamento vengono comunicate praticamente le stesse informazioni descritte sopra per i dipendenti pubblici, seppur con alcune piccole varianti. Tra queste la più importante ai fini dell'analisi è la tipologia di pagamento (diretto, anticipato). Infine in corrispondenza del datore di lavoro privato viene trasmesso il solo dato relativo alla matricola aziendale.

5 Inteso come sommatoria degli eventi di malattia che si sono verificati nell'anno solare in riferimento alla stessa unità di rilevazione.

11.3.3 Archivio INPS prestazioni dirette non pensionistiche

Come detto in precedenza l'archivio in oggetto permette di rilevare il complesso dei trattamenti di malattia (punti 1 e 2) pagati dall'Ente previdenziale direttamente ai lavoratori. La fonte amministrativa assume la classica struttura a matrice di dati; dove in riga figurano i beneficiari (identificati dal rispettivo codice fiscale) e in colonna le variabili: tipologia di trattamento e importo annuo associato. Nel caso in cui l'Ente provvedesse, nell'anno solare, ad effettuare più pagamenti in favore di un medesimo beneficiario in relazione a uno stesso titolo (trattamento), il dato finale riportato nella colonna importi incorporerebbe la sommatoria di tutti i versamenti.

11.3.4 Archivio INPS Emens delle differenze di accredito

L'archivio in questione rappresenta un segmento (o tabella) delle dichiarazioni Uniemens dell'INPS. Quest'ultima fonte madre, raccoglie le denunce aziendali relative alle posizioni lavorative dei propri dipendenti ed è strutturata in diversi moduli (sezioni), con contenuti informativi diversi quali: le caratteristiche demografiche e professionali (nome, cognome, matricola, codice fiscale, tipologia di qualifica, tipo di contribuzione) di ciascun lavoratore e le posizioni retributive e contributive ad esso associate. In particolare, la tabella EMENS_DIFFACCRE (differenza/accredito), include tra le varie informazioni fornite dall'azienda, il valore retributivo del lavoratore da considerare, nelle settimane in cui è avvenuto l'evento assicurato, a titolo di contribuzione figurativa. Tale valore denota la "retribuzione persa" dal lavoratore in conseguenza dell'evento malattia. In termini equivalenti, essa rappresenta l'ammontare della retribuzione che sarebbe stata considerata ai fini della contribuzione se il lavoratore avesse svolto la sua attività lavorativa regolarmente durante il periodo in cui è stato assente per l'evento assicurato[6]. Nel prossimo paragrafo si analizzerà il contenuto informativo e si evidenzieranno i limiti per il suo utilizzo ai fini della stima diretta del trattamento di cui al punto 7 precedentemente illustrato.

11.3.5 Archivio INPS Emens dei conguagli

Anche questo archivio rappresenta un segmento delle dichiarazioni Uniemens e include al suo interno, oltre alle informazioni base del dipendente e dell'azienda presso cui lavora, gli importi mensili messi a conguaglio dal datore, distintamente per le varie tipologie di evento assicurato. In particolare, tra queste ultime Figura l'indennità di malattia contrassegnata con l'etichetta: "CO_CNTR_INDENNIMAL" a cui è associato l'importo anticipato dall'azienda per conto dell'INPS. Come vedremo in seguito, la fonte INPS sui conguagli, diversamente dalle differenze di accredito, fornisce la base corretta per la ricostruzione degli importi (stima diretta) del trattamento anticipato di cui al punto 7.

11.3.6 Casellario INPS delle posizioni previdenziali attive

Il Casellario rappresenta, all'interno sistema pensionistico pubblico italiano, l'anagrafe generale delle posizioni assicurative dei lavoratori iscritti all'Assicurazione Generale Obbli-

6 Per ulteriori approfondimenti si rinvia al Capitolo 10 redatto da Cirelli e Consolini.

gatoria o forme sostitutive. Istituito presso l'INPS, con L.243 del 23 agosto 2004 art. 1 c.23, è deputato alla raccolta, conservazione e gestione dei dati relativi ai lavoratori iscritti: a) all'assicurazione generale obbligatoria per l'invalidità, la vecchiaia e i superstiti dei lavoratori dipendenti, anche con riferimento ai periodi di fruizione di trattamenti di disoccupazione o di altre indennità o sussidi che prevedano una contribuzione figurativa; b) ai regimi obbligatori di previdenza sostitutivi dell'assicurazione generale obbligatoria per l'invalidità, la vecchiaia e i superstiti o che ne comportino comunque l'esclusione o l'esonero; c) ai regimi pensionistici obbligatori dei lavoratori autonomi, dei liberi professionisti e dei lavoratori di cui all'articolo 2, comma 26, della legge 8 agosto 1995, n. 335; d) a qualunque altro regime previdenziale a carattere obbligatorio; e) ai regimi facoltativi gestiti dagli enti previdenziali. Ai fini della sperimentazione esso è utilizzato per la rilevazione dei periodi di contribuzione figurativa, distinti per evento, e delle retribuzioni ad esse associate. Si sovrappone, come contenuto informativo, all'archivio del precedente punto (EMENS_DIFFACCRE).

11.4 Malattia: consolidamento archivi, criteri di calcolo e disegno multi-fonte

L'approccio seguito nella sperimentazione riprende per sommi capi l'impostazione adottata nel precedente capitolo 10, dove l'obiettivo era lo scorporo della Cassintegrazione (Cigo, Cigs, Sol) anticipata dai datori rispetto alla retribuzione degli stessi dipendenti del privato. Tuttavia, in virtù del maggior dettaglio informativo di cui dispone, la ricerca si differenzia dalla precedente, per la diversa scelta delle basi dati e per l'applicazione di un innovativo disegno multi-fonte, dove alla stima diretta della prestazione/trattamento si affianca quella indiretta.

Preliminare all'implementazione dell'approccio indiretto è il consolidamento della fonte del dato primaria: Certificazioni telematiche della malattia dei dipendenti del settore privato/pubblico. L'ostacolo principale al raggiungimento di una corretta (o almeno il più possibile scevra da errori) stima sul numero dei giorni di malattia, da associare a ciascun lavoratore, è la ricostruzione dei periodi di malattia. L'evento malattia, per sua natura, può verificarsi o meno in modo ricorrente durante l'anno. L'operazione di consolidamento, come vedremo, mostra diversi gradi di complessità, poiché i dati contengono in una certa misura errori di digitazione, duplicazione e sovrapposizione. Infine, ma non ultimo, vi è l'obiettivo dell'individuazione della cosiddetta carenza (ovvero un sottoinsieme del periodo di malattia) da tenere distinta dai restanti giorni, in quanto essa prefigura un regime di erogazione di natura privata, diverso rispetto al circuito assicurativo-previdenziale (regime pubblico).

La base dati utilizzata per la sperimentazione è costituita dall'insieme delle persone fisiche appartenenti alle liste campionarie delle indagini IT-SILC 2014-2015 (d'ora in poi: IT-SILC14_15), per le quale era possibile l'abbinamento di un codice fiscale. L'unione dei campioni teorici IT-SILC14_15 è formato da 92.754 unità (soggetti), di esse 91.840 sono in possesso di un codice fiscale validato dalla Sogei (99,0%).

11.4.1 La stima indiretta della prestazione di malattia e l'elemento base: periodo

Il database INPS relativo alle Certificazioni telematiche della malattia dei dipendenti del privato/pubblico rappresenta, come già detto, la fonte primaria alla base della stima indiretta del trattamento di malattia, ovvero il numero dei giorni indennizzati.

In particolare, quello riferito al settore privato contiene 6.462 soggetti (codici fiscali) distinti che trovano un abbinamento con la lista campionaria di partenza (7%). In totale sono inclusi 18.058 record che, una volta ripuliti dalle duplicazioni, si riducono a 17.885 record. La chiave primaria utilizzata per il consolidamento è data dalla combinazione delle variabili base (chiave multipla): <Codice_fiscale||Matricola_aziendale||Inizio_PR||Fine_PR>. L'errore di digitazione che si presenta con più frequenza coinvolge la variabile "data fine prognosi (Fine_PR)", dove figurano una serie di valori "0" nella relativa struttura temporale Anno-Mese-Giorno (AAAA.MM.GG). L'ipotesi utilizzata per la correzione del valore inammissibile è di farlo coincidere con la data di inizio periodo di malattia (Inizio_PR).

Figura 11.2 - Prestazione monetaria non pensionistica e componenti: indennità di malattia

	codfi	INIZIO_PR	FINE_PR	DATA_RIL	FLAG_EVEN	AZ_M	FLAG_INF	FLAG_PAG	QUALIF	SETTOR
490	BLT	20150902	00000000	20150902	I	49374		A	01	1
1066	BRN	20150417	00000000	20150417	I	13167		A	I2	2
1080	BRN	20151023	00000000	20151023	I	49683		A	02	2
1582	BSS	20150520	00000000	20150520	I	13111		A	I2	2
1889	CCC	20150419	00000000	20150419	I	49370		A	I2	2
2653	CND	20150504	00000000	20150505	I	54023		A		
3160	CRB	20150925	00000000	20150925	I	49710		A	02	2
3456	CRR	20150114	00000000	20150114	I	15066		A	02	2
3629	CRT	20150316	00000000	20150316	I	34133		A	01	1
3732	CSG	20150318	00000000	20150318	I	24082		A	01	1
4852	DMR	20150508	00000000	20150508	I	13127		A	I2	2
4918	DNC	20151210	00000000	20151210	I	85028		A	02	2
5070	DPR	20150209	00000000	20150209	I	81369		A	01	1
5464	FDD	20151123	00000000	20151123	I	15084		A	01	1

Per il singolo evento di malattia, il passaggio dal periodo da calendario (Data_Inizio-Data_Fine) al numero di giorni è un'operazione semplice se il periodo ricade in uno stesso mese, mentre richiede una trasformazione (moltiplicatori) quando il periodo è a cavallo di più mesi. Nel caso di superamento dell'anno solare si utilizza come data convenzionale di fine trattamento il 31 dicembre dell'anno base. Come si evince da questa introduzione il periodo di malattia è caratterizzato da due elementi: *Periodo_malattia=(data_inizio, data_fine)*. Il calcolo dei giorni di malattia può essere ottenuto tramite la differenza della "data_fine (Fine_PR)" e "data_inzio (Inizio_PR)" periodo opportunamente trasformate (aggiungendo un'unità). Il presentarsi di duplicazioni relative a stessi periodi di malattia su medesimi lavoratori, con associati differenti datori di lavoro, suggerisce il collassamento della chiave primaria, tramite l'eliminazione dell'elemento datore. L'uso della nuova chiave ridotta <Codice_Fiscale||Inizio_PR||Fine_PR> genera 17.681 record distinti (Tavola 11.1). Dalla seguente Tavola 11.1 si ricavano interessanti spunti per l'analisi: nel campione vi sono 2.709 percettori di indennità di malattia per i quali si presenta un solo evento di malattia all'anno (41,9%), mentre per i rimanenti 3.753 beneficiari (58,1%) lo stesso evento mostra delle ricorrenze (nel 23,5% dei casi si tratta di periodi registrati in due diverse dichiarazioni).

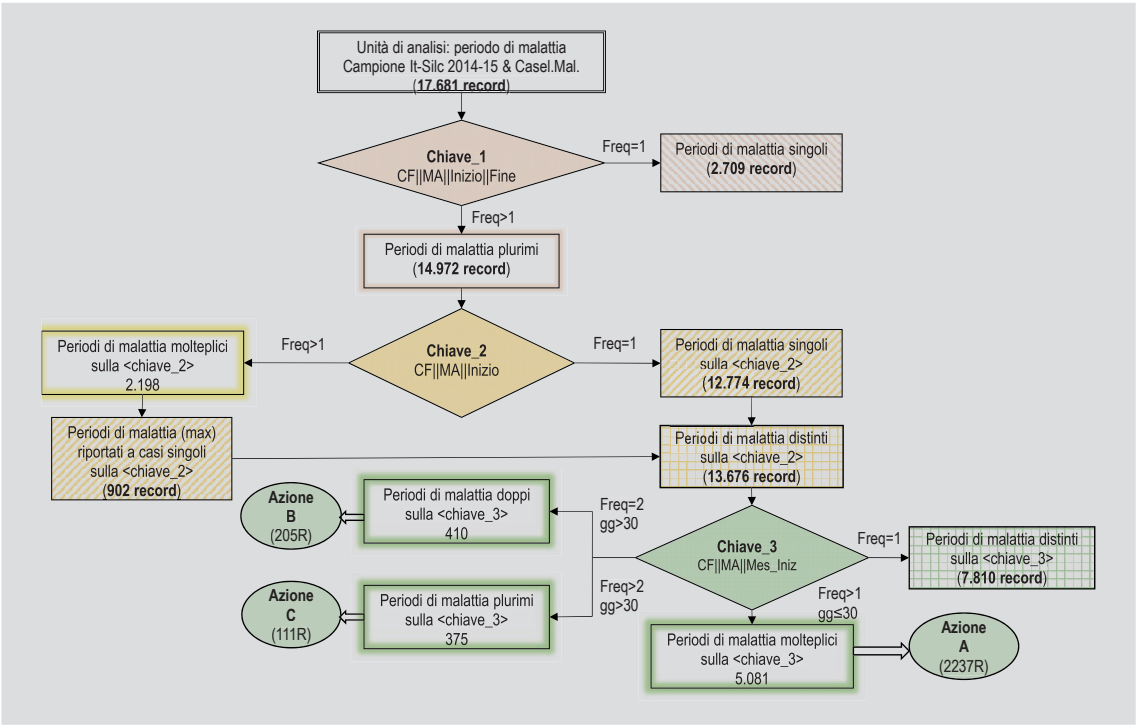
Tavola 11.1 - Distribuzione di frequenza dei periodi malattia per titolare: dipendenti del privato (valori percentuali)

Molteplicità periodi	Frequenza codici fiscali	%	Codici fiscali	% Cumulata codici fiscali	Frequenza periodi
1	2.709		41,92	41,92	2.709
2	1.519		23,51	65,43	3.038
3	818		12,66	78,09	2.454
4	456		7,06	85,14	1.824
5	293		4,53	89,68	1.465
6	180		2,79	92,46	1.080
7	101		1,56	94,03	707
8	110		1,7	95,73	880
9	78		1,21	96,94	702
10+	198		3,11	100,00	2.822
Totale	**6.462**		**100,00**	**-**	**17.681**

Per identificare i periodi di malattia distinti da quelli sovrapposti, cioè se si tratta di una ricorrenza vera o apparente, sono stati introdotti alcuni criteri di consolidamento. Nei dati in esame si osservano alcune regolarità, per cui accade che un periodo possa includere il successivo o i successivi, sottintendendo il fatto che si tratti di una dichiarazione sostituiva delle altre. Il problema che un analista dei dati amministrativi si trova spesso di fronte è quello di individuare l'origine che ha dato luogo alla molteplicità delle dichiarazioni: se è riferibile a una integrazione o una sostituzione della precedente. A seconda della natura della dichiarazione, si adottano criteri differenti per il consolidamento delle relative informazioni.

Il problema della sovrapposizione dei periodi di malattia in capo a uno stesso lavoratore è stato affrontato, inizialmente, mediante il raggruppamento e ordinamento dei dati in base alla seguente chiave: <Codice_Fiscale||Matricola_Aziendale||Inizio_PR>. Di conseguenza il collettivo di partenza è stato suddiviso in due sottoinsiemi distinti: 1) periodi iniziali (Inizio_PR) singoli, cioè con frequenza unitaria; 2) periodi iniziali molteplici, ovvero con frequenza superiore all'unità. In relazione alla prima casistica non si riscontra alcun problema di sovrapposizione

Figura 11.3 - Albero decisionale per la ricomposizione delle sovrapposizioni dei periodi di malattia



sul primo elemento del periodo (Inizio), mentre ciò non vale per i secondi (figura 11.3). In relazione a questi ultimi si ipotizza la presenza di dichiarazioni sostitutive; pertanto si seleziona, tra i tanti, il record corrispondente al periodo più ampio (ovvero quello con la data di fine periodo più lontana nel tempo).

Come si evince a sinistra della Figura 11.3, tale operazione comporta la ricerca del valore massimo tra le date di fine periodo, e riduce la dimensione iniziale del segmento analizzato da 2.198 a 902 record distinti.

La successiva fase consiste nella ricomposizione dei periodi di malattia la cui data iniziale ricade nello stesso mese (Chiave_3 di Figura 11.3). Nella definizione dei criteri, si è seguito un approccio misto logico-deduttivo ed euristico, cioè sulla base delle regolarità presenti nella struttura dei dati. Ad esempio, l'evidenza empirica mostra che si manifestano raramente sovrapposizioni quando la somma dei giorni di malattia per i vari periodi (con stesso inizio mese) risulta inferiore a 30. Quindi l'azione meno esposta a errori (A di Figura 11.3) consiste nella somma delle giornate per le varie dichiarazioni. Questo aspetto è intuitivo poiché, più sono brevi periodi di malattia, minore è il rischio di sovrapposizioni. Diverso è il discorso quando i periodi sono più dilatati (>30 giorni): le casistiche fortunatamente sono meno frequenti. In fase di correzione abbiamo comunque voluto applicare una partizione, separando i record doppi (molteplicità 2) da quelli plurimi (convenzionalmente così chiamati per definire le molteplicità 3+). Il motivo di tale distinguo risiede nel fatto che nel caso di record doppi la quadratura è più semplice e immediata. Senza voler scendere troppo nei dettagli tecnici, le azioni finali consistono nello scegliere il periodo più lungo (in caso di sovrapposizione), la somma dei periodi in caso di separazione, o il periodo più breve in presenza di errori.

Le procedure descritte in Figura 11.3 non esauriscono l'insieme delle operazioni di data cleaning finalizzate all'identificazione e correzione di errori nella struttura temporale delle dichiarazioni di malattia. Infatti, nella fase di trattamento dati si è tenuto in debita considerazione la presenza di incoerenze dovute alla sovrapposizione di periodi, anche in situazioni dove fossero differenti i mesi di inizio evento. Inoltre più in generale, si è proceduto a eliminare nel conteggio i giorni ripetuti a cavallo di due periodi di malattia contigui, utilizzando la funzione *LAG* (ritardo) del software statistico SAS. Grazie a questa funzione si è potuto confrontare la data di inizio del nuovo periodo con quella di fine evento del precedente. L'output finale assume la configurazione illustrata in Figura 11.4. All'interno della struttura dati sono presenti tutte le variabili utili ai fini dell'analisi.

Figura 11.4 - Frame del database di output sui giorni di malattia concessi ai lavoratori del privato

	codfis	NR_RECC	NR_REC	GG_MAL	GG_CAREN	MES_INIZ	MES_FINE	GG_MAL_DAT	AZ_M	GG_MAL_DATOR2
1	BAUB	1	3	10	8	01	07	10	9300	.
2	BAUD	1	1	6	3	01	01	6	9107	.
3	BAUF	1	4	12	11	02	11	12	9110	.
4	BBRR	1	1	4	3	04	04	4	5202	.
5	BBRS	1	2	5	5	01	02	5	6803	.
6	BBTD	1	1	5	3	05	05	5	8112	.
7	BBTM	1	1	2	2	03	03	2	6506	.
8	BBTM	1	1	5	3	07	07	5	0917	.
9	BCCB	1	1	17	3	10	11	17	1316	.
10	BCCC	1	1	17	3	11	12	17	5602	.
11	BCCC	1	1	6	3	04	04	6	2304	.
12	BCCC	1	2	15	6	03	09	15	5411	.
13	BCCD	1	1	10	3	05	05	10	8134	.
14	BCCD	1	2	13	3	08	09	13	0605	.

Oltre ai giorni di malattia complessivi nell'anno solare (*GG_MAL*), figurano anche la somma dei giorni di carenza associati ai distinti periodi (*GG_CAREN*), nonché il mese in cui ricade l'inizio del periodo del primo evento (*MES_INIZ*) e il mese in cui termina l'ultimo (o anche unico) evento di malattia (*MES_FINE*). In presenza di più datori di lavoro si è ulteriormente tenuto conto della ripartizione dei giorni per datore. Per necessità di sintesi figurano i soli primi due datori, d'altronde è molto raro il caso di compresenza di tre o più datori nell'anno.

Di seguito si riportano i valori di sintesi della distribuzione dei giorni di malattia per i dipendenti del privato (Tavola 11.2). Nel complesso, nel settore privato figurano 6.462 di percettori di indennità di malattia appartenenti al campione IT-SILC14_15, a cui corrisponde un totale di 110.246 giorni di assenza retribuita nell'anno solare, per una media annua di 17,1 giorni.

Tavola 11.2 - Valori di sintesi della distribuzione dei giorni malattia (GG_MAL): dipendenti del privato

Percettori	Somma	Media	Minimo	1° Quartile	Mediana	3° Quartile	Massimo
6.462	110.246	17,06	1	4	7	17	288

Applicando lo stesso approccio alla banca dati INPS sulle Certificazioni telematiche dei dipendenti pubblici, si individuano 3.074 percettori di indennità di malattia (Tavola 11.3) con associate 9.834 dichiarazioni distinte di malattia, cioè ripulite dalle duplicazioni. I titolari di indennità nel pubblico impiego, osservati nel campione, sono poco meno della metà (47,5%) di quelli appartenenti al settore privato.

Tavola 11.3 - Distribuzione di frequenza dei periodi malattia per titolare: dipendenti pubbliciali)

Molteplicità periodi	Frequenza codici fiscali	% codici fiscali	% cumulata codici fiscali	Frequenza periodi
1	1.072	34,9	34,9	1.072
2	683	22,2	57,1	1.366
3	445	14,5	71,6	1.335
4	266	8,7	80,2	1.064
5	171	5,6	85,8	855
6	151	4,9	90,7	906
7	66	2,2	92,8	462
8	59	1,9	94,8	472
9	35	1,1	95,9	315
10+	126	4,1	100,0	1.987
Totale	**3.074**	**100,0**	**-**	**9.834**

Gli stessi titolari usufruiscono di 55.711 giorni complessivi di assenza retribuita nell'anno solare, per una media annua di poco oltre 18 giorni, un giorno in più del settore privato.

Tavola 11.4 - Valori di sintesi della distribuzione dei giorni malattia (GG_MAL): dipendenti pubblici

Percettori	Somma	Media	Minimo	1° Quartile	Mediana	3° Quartile	Massimo
3.074	55.601	18,09	1	4	7	16	328

11.4.2 La stima diretta della prestazione di malattia: importo erogato

Per quanto concerne la stima diretta del trattamento individuale INPS che si incarica del versamento all'assicurato (vedasi archivio prestazioni non pensionistiche INPS di cui al §11.3.3), non sussiste alcuna difficoltà di calcolo.

Figura 11.5 - Frame del database INPS sulle prestazioni (dirette) non pensionistiche

	COD	CODICE_TRATTAMENTO	IMPORTO_TRATT_ANNO_CORR	IMPORTO_TRATT_ANNO_PREC
1	BGL	F01	0.00	154.78
2	BLD	F01	222.36	0.00
3	BLH	F01	0.00	530.67
4	BNA	F01	0.00	1112.11
5	BNL	F01	210.81	0.00
6	BN	F01	1850.79	0.00
7	BNV	F01	1585.58	0.00
8	BRB	F01	86.87	0.00
9	BRC	F01	0.00	302.50
10	BRC	F01	64.94	0.00
11	BRR	F01	862.34	0.00
12	BRT	F01	2137.59	0.00
13	BRT	F01	95.15	0.00
14	BRT	F01	427.00	0.00

Una volta selezionato il titolo ad essa riferita: trattamento "F01" (malattia) e trattamento "N01" (malattia e idoneità ex-Ipsema); si procede a sommare i valori versati ai vari beneficiari (COD) nell'anno di riferimento. La precedente Figura 11.5 illustra la struttura del database in esame.

Dalla seguente Tavola 11.5 si evince lo scarso peso del numero di beneficiari e degli importi associati, soprattutto se messo a confronto col collettivo dei soggetti inclusi nell'archivio telematico sui certificati di malattia dei dipendenti privati. I titolari di prestazione a erogazione diretta costituiscono appena il 3% di tutto il settore privato.

Tavola 11.5 - Distribuzione di frequenza dei periodi malattia per titolare: dipendenti pubbliciali)

TIPOLOLOGIA	Percettori	Somma	Media	1° Quartile	Mediana	3° Quartile
Malattia generica	176	131.686	748	101	428	1.078
Malattia ex Ipsema	18	95.492	5.305	1.050	5.467	8.298
Totale	**194**	**227.178**	**1.171**	**123**	**488**	**1.274**

La stima degli importi erogati al restante 97% del settore privato rimanda all'esplorazione di fonti del dato alternative. Tra queste, si è scelto in partenza l'archivio EMENS_DIFFACCRE dell'INPS, già utilizzato con successo nella stima delle prestazioni legate alla cassintegrazione. Ai fini del suo possibile impiego, è stata valutata la corrispondenza con gli eventi registrati nell'archivio telematico delle certificazioni di malattia. Da una prima

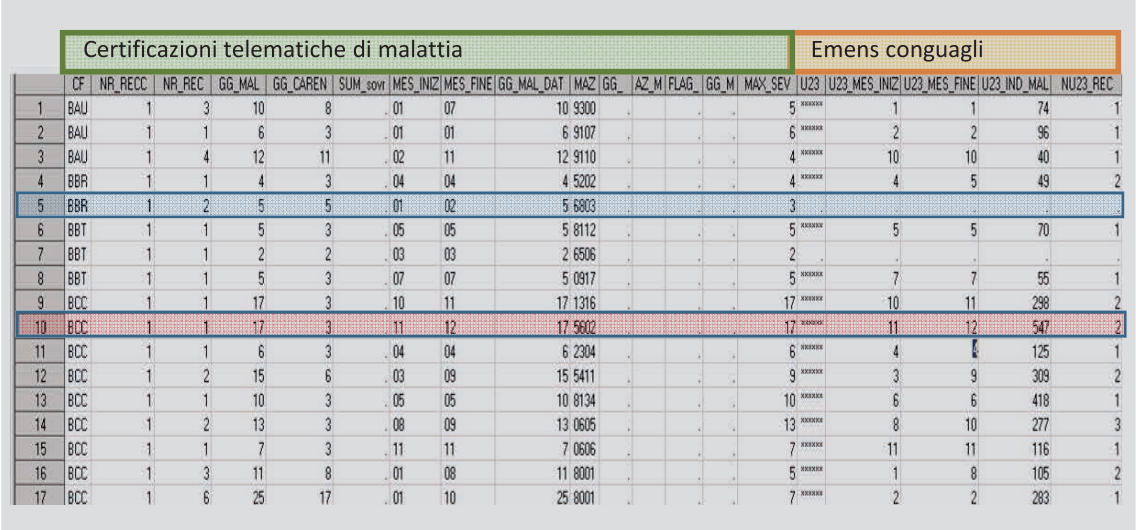
ispezione visiva emerge tuttavia il suo limite principale: la restrizione del campo di osservazione ai soli eventi di durata non inferiore a sette giorni e la conseguente mancata rilevazione di quei periodi brevi di malattia, in parte a carico del datore di lavoro (primi tre giorni di carenza) e in parte dell'INPS (dal 4° al 6° giorno). La Tavola 11.6 mostra un'analisi comparata della copertura della fonte *Emens_Diffacr* rispetto alla fonte *Emens_Conguagli* in relazione ai vari periodi di malattia. Sono esclusi dal campo di osservazione quegli eventi associati all'erogazione diretta INPS e all'intersezione dei collettivi di dipendenti del settore privato e pubblico.

Tavola 11.6 - Distribuzione di frequenza periodi malattia per la classe di ampiezza, numero eventi, numero dichiarazioni Emens Diffacr e Emens Conguagli

Classe di ampiezza evento (in giorni)	Numero eventi	Numero dichiarazioni Emens Diffaccr	% copertura Emens Diffaccr	Numero dichiarazioni Emens Conguagli
1-3	1.843	27	1,5	66
4-6	1.974	106	5,4	1.758
7+	2.293	2.063	90,0	2.167
Totale	**6.101**	**2.196**	**35,9**	**3.991**

Come anticipato, la fonte *Emens_Diffacr* raggiunge una bassa copertura (5,4%) per gli eventi di malattia compresi nell'intervallo 4-6 giorni. Tale lacuna ci induce a dover perseguire altre strade. L'orientamento della ricerca si sposta verso l'altro candidato, il database *Emens_Conguagli*, la cui copertura nella stessa classe 4-6 giorni è di poco inferiore al 90%. Ai fini della stima diretta, la fonte sugli importi a conguaglio rappresenta la base ideale; in quanto, in linea di principio, dovrebbe fotografare l'istante in cui gli importi sono anticipati in busta paga (mensilmente) senza significative differenze temporali rispetto al momento della loro registrazione (per scopi autorizzativi da parte dell'Istituto previdenziale). Per saggiare tale ipotesi si è resa necessaria la verifica sul suo allineamento temporale rispetto all'archivio delle certificazioni telematiche di malattia, il quale fotografa il momento in cui si è manifestato l'evento morboso. In Figura 11.6 vengono presentati gli elementi principali dell'output che integra le due fonti tramite il codice fiscale degli stessi titolari (chiave di abbinamento). Il record evidenziato in azzurro a posizione # 5 (come pure il successivo #7), mostra un esempio di consistenza tra il dato sulle dichiarazioni a conguaglio (importi) e quello delle certificazioni telematiche di malattia (periodi). Infatti, l'assenza di dati relativi agli importi da conguagliare si spiega con la completa capienza del periodo di carenza (GG_CAREN=5) rispetto al totale dei giorni di malattia (GG_MAL=5), ovvero tutti i giorni di malattia sono a totale carico del datore. Interessante è l'analisi del contenuto relativo al record #10 (evidenziato in rosso) che sarà preso come caso studio per la stima diretta.

Figura 11.6 - Frame del database integrato "Emens conguagli" e "Certificazioni telematiche di malattia"



	Certificazioni telematiche di malattia													Emens conguagli				
	CF	NR_RECC	NR_REC	GG_MAL	GG_CAREN	SUM_sovr	MES_INIZ	MES_FINE	GG_MAL_DAT	MAZ	GG_	AZ_M	FLAG_	GG_M	MAX_SEV	U23	U23_MES_INIZ	U23_MES_FINE	U23_IND_MAL	NU23_REC
1	BAU	1	3	10	8	.	01	07	10 9300	.	.	.	.		5	******	1	1	74	1
2	BAU	1	1	6	3	.	01	01	6 9107	.	.	.	.		6	******	2	2	96	1
3	BAU	1	4	12	11	.	02	11	12 9110	.	.	.	.		4	******	10	10	40	1
4	BBR	1	1	4	3	.	04	04	4 5202	.	.	.	.		4	******	4	5	49	2
5	BBR	1	2	5	5	.	01	02	5 6803	.	.	.	.		3	.	.	.	.	.
6	BBT	1	1	5	3	.	05	05	5 8112	.	.	.	.		5	******	5	5	70	1
7	BBT	1	1	2	2	.	03	03	2 6506	.	.	.	.		2	.	.	.	.	.
8	BBT	1	1	5	3	.	07	07	5 0917	.	.	.	.		5	******	7	7	55	1
9	BCC	1	1	17	3	.	10	11	17 1316	.	.	.	.		17	******	10	11	298	2
10	BCC	1	1	17	3	.	11	12	17 5602	.	.	.	.		17	******	11	12	547	2
11	BCC	1	1	6	3	.	04	04	6 2304	.	.	.	.		6	******	4	4	125	1
12	BCC	1	2	15	6	.	03	09	15 5411	.	.	.	.		9	******	3	9	309	2
13	BCC	1	1	10	3	.	05	05	10 8134	.	.	.	.		10	******	6	6	418	1
14	BCC	1	2	13	3	.	08	09	13 0605	.	.	.	.		13	******	8	10	277	3
15	BCC	1	1	7	3	.	11	11	7 0606	.	.	.	.		7	******	11	11	116	1
16	BCC	1	3	11	8	.	01	08	11 8001	.	.	.	.		5	******	1	8	105	2
17	BCC	1	6	25	17	.	01	10	25 8001	.	.	.	.		7	******	2	2	283	1

Il numero totale di giorni di malattia ad esso riferiti è pari a 17, di cui 3 per la carenza, mentre l'importo conguagliato per lo stesso evento è uguale a 547 euro.

Relativamente allo stesso record, si nota un perfetto allineamento delle informazioni sui mesi in cui sono verificati gli eventi (MES_INIZ=11; MES_FINE=12) e quelli in cui è stata inoltrata la richiesta di conguaglio all'INPS (U23_mes_iniz=11; U23_mes_fine=12).

La Tavola 11.7 mostra la distribuzione di frequenza degli eventi a seconda dell'esito derivante dal confronto delle due fonti integrate.

Non sempre è verificata la perfetta coincidenza dei mesi, in virtù del fatto che l'insieme degli eventi, i cui elementi sono i distinti periodi di malattia, è più esteso rispetto al secondo, costituito dalle singole dichiarazioni di conguaglio (solo per periodi superiori al terzo giorno). Inoltre, essendo data facoltà al datore di effettuare la dichiarazione il mese successivo al verificarsi dell'evento, sussistono in una percentuale ridotta di casi dei disallineamenti di un mese. Si rigetta la casistica che presenta incompatibilità col periodo di malattia da certificazione telematica e il cui mese di inizio erogazione è gennaio con termine non oltre marzo (ricollegabile al precedente anno).

Tavola 11.7 - Distribuzione di frequenza degli eventi malattia per esito del confronto delle fonti INPS

Esiti confronto mesi di erogazione conguagli vs mesi evento malattia Archivio Certificazioni Telematiche di malattia (CTM)	Numero eventi	%	Azione conseguente
Mese/i di erogazione compatibile/i con CTM	3.650	91,4	accettazione
- di cui mese_inizio e mese_fine coincidenti con CTM	*(2.125)*	*(52,2)*	*accettazione*
Mese_inizio erogazione successivo a mese_fine CTM	154	3,9	accettazione
Mese_inizio erogazione diverso da gennaio e incom-patibile con CTM	83	2,1	accettazione
Erogazione: Mese_inizio gennaio e Mese_fine non oltre marzo incompatibile con CTM	23	0,6	rifiuto
Erogazione: Mese_inizio gennaio e Mese_fine da aprile i poi e incompatibile con CTM	81	2,0	accettazione
Totale	**3.991**	**100**	**-**

Prendendo a titolo dimostrativo le informazioni del record #10, si può delineare lo schema del procedimento di stima degli importi nel suo complesso. Il primo passo consiste nel calcolo della retribuzione media giornaliera (*RMG*) implicito nella commisurazione dell'indennità malattia. Occorre puntualizzare che il calcolo dell'indennità

di malattia contiene diversi elementi di complessità, poiché differenziata in base alle caratteristiche professionali, contrattuali e alla durata dell'evento (ad esempio: se impiegato, operaio, …, o a seconda della tipologia di contratto collettivo nazionale, di retribuzione oraria o mensilizzata, eccetera). Di seguito verrà formalizzato il procedimento di calcolo degli importi:

$IMP\_MAL\_ANTIC^{L} = CO\_CNTR\_INDENNIMAL$[7] [€ 547];
$IMP\_MAL\_ANTIC^{F}=(IMP\_MAL\_ANTIC^{L} \div Coef.Lord.)$[8] [547÷1,1012=€ 496,7];
$RMG=(IMP\_MAL\_ANTIC^{L} \div (GG\_MAL - GG\_CAREN)\cdot \%IND)$[9] [(€ 547÷7)= € 78,1];
$RMG^{F}= (RMG \div Coef.Lord.)$ [€ 71];
$IMP\_CAREN^{F} = (GG\_CAREN \cdot RMG^{F})$ [(3·€ 71)=€ 213];
$IND\_MAL\_TOT^{F}= IMP\_CAREN^{F}+IMP\_MAL\_ANTIC^{F}$ [(€ 213+€ 496,7)= € 609,7].

Il termine "IMP_MAL_ANTIC" è riferito all'importo corrisposto al lavoratore e anticipato dal datore (incluso nella fonte *EMENS_CONGUAGLI*); la variabile "*RMG*" denota a sua volta la retribuzione lorda media giornaliera. Da questi semplici passaggi si deduce che l'indennità totale di malattia nel privato è formata da due distinti elementi (trattamenti), rispettivamente dati dalla carenza erogata dal datore (IMP_CAREN) e dall'indennità di malattia anticipata dal datore per conto dell'INPS (IMP_MAL_ANTIC). Nello schema precedente l'apice (^{F}) contraddistingue una componente economica riferita all'imponibile fiscale e non già all'imponibile previdenziale (come è il caso della *RMG*). Infatti, ai fini del calcolo dell'indennità di malattia al netto del prelievo fiscale e contributivo, obiettivo ultimo della stima diretta per gli scopi d'indagine, è più semplice rapportarsi alla componente lorda fiscale. Occorre precisare che il calcolo della carenza si fonda, in realtà, su un approccio indiretto; cioè sulla base del numero di giorni di malattia moltiplicato per la retribuzione media giornaliera (componente lorda fiscale). Pertanto ai fini dell'analisi, nella stima globale del fenomeno si ricorre ad un approccio "misto": diretto (malattia anticipata) e indiretto (periodo di carenza).

Lo stesso procedimento indiretto si applica, in via esclusiva, al caso dei dipendenti pubblici; la stima finale è data dal prodotto tra queste due grandezze. Sempre in riferimento al pubblico impiego, la retribuzione media giornaliera si desume poi da un'altra fonte non citata (dichiarazioni fiscali presenti in CU) e si ricava dal rapporto tra reddito imponibile annuo da lavoro dipendente e il numero di giorni lavorati.

In sintesi il ricorso a un disegno multi-fonte, come quello mostrato nel presente paragrafo, costituisce l'unica soluzione metodologica quando ci si trova di fronte alla necessità di ricostruire un puzzle, alla cui origine vi è la frammentazione delle informazioni (elementi di calcolo) diffuse su una moltitudine di archivi e, al tempo stesso, la complessità del fenomeno per motivi normativo-istituzionali. Nel caso della prestazione in esame, cioè l'indennità giornaliera di malattia, figurano ben otto sottolivelli altrimenti detti trattamenti, per la

7 Così indicata nel database *EMENS_CONGUAGLI*, mentre in Figura 11.6 compare sotto l'etichetta: "U23_IND_MAL".

8 Il coefficiente di lordizzazione varia a seconda dell'aliquota contributiva a carico del lavoratore: se quest'ultima è pari al 9,19% corrisponde a {1,1012}, mentre se l'aliquota è commisurata al 9,49% è uguale a {1,10485}.

9 Il termine "%IND" sta ad indicare la percentuale di indennizzo, solitamente fissata al 50%, per cui il numero di giorni di malattia a carico INPS va dimezzato per riportarlo su base unitaria. Nel caso degli impiegati del terziario, la stessa percentuale a partire dal 21° giorno cresce al 66%. Mentre per tutte le figure professionali si azzera quando il totale dei giorni di malattia nell'anno solare supera il valore soglia 180, cioè il cosiddetto periodo di comporto (in genere per i malati oncologici e per le persone con complicazioni post-operatorie sono previste forme di tutela disciplinate dai CCNL).

cui rilevazione sono state individuate sei fonti del dato. Tra l'altro, queste ultime risultano non essere esaustive e se ne deve aggiungere una settima: Certificazione Unica (CU) dell'Agenzia delle Entrate.

11.5. Analisi delle incoerenze, riconciliazione tra fonti e validazione esterna

Una volta definiti i criteri per la stima degli importi, il passo successivo consiste nell'individuazione delle incoerenze presenti nei dati, mediante il processo validazione tra le fonti utili allo scopo. Nel nostro caso, l'applicazione delle classiche procedure di *data cleaning* (*errors identification and removal*) sulle incoerenze interne[10] non è sufficiente a garantire la qualità dell'informazione utilizzata per la stima del fenomeno. Per assicurare una migliore accuratezza del dato, si farà ricorso alla riconciliazione dei microdati, la cui fase preliminare consiste nella definizione delle relazioni che legano le variabili oggetto di studio lungo tutte le fonti del dato (European Commission, 2014; Pannekoek, 2011; van der Laan, 2000). Nella sperimentazione del calcolo dell'indennità di malattia occorre dunque verificare la consistenza dei dati e le relazioni che vincolano: giorni di malattia, periodo di carenza, regole alla base dell'erogazione della prestazione, nonché i valori degli importi conguagliati dal datore e dichiarati nell'Uniemens.

Si denoti con **Y** la matrice dati rettangolare ($N \times 3$), avente N-righe (soggetti) e 3 colonne (indicanti nell'ordine le variabili oggetto di studio: giorni di carenza, giorni di malattia e dichiarazioni a conguaglio). In particolare, si definisca con $y_i=(y_{i1},y_{i2},m_{i3})$ l'i-esima riga della stessa matrice, il cui elemento y_{ij} è il valore assunto dalla variabile Y_j in corrispondenza all'i-esimo soggetto; mentre $m_{i3}=f(y_{i3})$ rappresenta l'indicatore di valore missing corrispondente alla variabile M_3, tale che $m_{i3}=1$ se y_{i3} è mancante sulla variabile Y_3 e $m_{i3}=0$ se y_{i3} è presente.

Le informazioni contenute nell'archivio integrato, illustrato nella precedente Figura 11.6, per garantire la coerenza devono soddisfare il seguente vincolo relazionale:

$$\forall_i := \{y_{i1} \in Y_1 ; y_{i2} \in Y_2 : 0 < y_{i1} = y_{i2}\} \Leftrightarrow \exists_i := \{m_{i3} \in M_3 : m_{i3} = 1\} \quad [1]$$

Dove y_{i1} indica il valore della variabile giorni di carenza (Y_1) per l'i-esimo individuo, mentre y_{i2} denota il valore della variabile giorni di malattia (Y_2) associato allo stessa persona, infine m_{i3} sottende l'assenza/presenza di una dichiarazione Emens a conguaglio $(M_3=f(Y_3))$ in capo allo stesso soggetto. In altri termini, in vincolo stabilisce che per qualunque unità la presenza di un periodo di carenza capiente (rispetto al numero totale di giorni di malattia) si deve associare una dichiarazione nulla sull'archivio dei conguagli e viceversa. Tale relazione risulta soddisfatta in 1.807 casi; mentre è violata su 51 individui, cioè è presente un importo a conguaglio nonostante il periodo di malattia coincida con la carenza ($x_i=y_i$). Qualora le incoerenze si riferiscano a dichiarazioni presentate a gennaio, si ipotizza che l'evento morboso sia iniziato (e saldato in busta paga) l'anno precedente; pertanto in 23 casi si è ritenuto opportuno escludere tanto importo a carico INPS quanto quello di del periodo di carenza (si veda Tavola 11.7). Nei restanti 28 casi si ammette la possibilità di un errore nell'archivio telematico delle malattie e quindi si include l'importo a conguaglio nella stima finale.

Il secondo vincolo che deve essere assicurato per garantire la coerenza dei dati è il seguente:

$$\forall_i := \{y_{i1} \in Y_1' ; y_{i2} \in Y_2' : 0 < y_{i1} < y_{i2}\} \Leftrightarrow \exists_i := \{m_{i3} \in M_3' : m_{i3} = 0\} \quad [2]$$

10 Intese come incoerenze individuate sulla base delle informazioni interne alla stessa fonte.

Dove le variabili Y_1' (giorni di carenza) e Y_2' (giorni di malattia) e M_3' (assenza/presenza di una dichiarazione a conguaglio) insistono su un sottoinsieme delle dichiarazioni telematiche dei dipendenti privati, con l'esclusione delle unità che hanno intersezione non nulla con l'archivio dei dipendenti pubblici (per definizione non assicurati all'INPS) e di quelle il cui periodo di malattia è associato a un pagamento diretto (e quindi non anticipato). Pertanto dal collettivo iniziale di 4.600 dichiarazioni telematiche riportanti giorni di malattia superiori alla carenza, andranno defalcate 283 unità che esulano dall'anticipo della prestazione. Il vincolo sopramenzionato stabilisce la regola che in presenza di una qualunque comunicazione telematica indicante almeno un giorno indennizzabile ($x_i<y_i$) vi deve essere una dichiarazione a conguaglio per lo stesso evento e viceversa. Il vincolo è soddisfatto in 3.940 casi[11]; mentre per 377 unità risulta violato, non essendovi alcuna dichiarazione a conguaglio su periodi indennizzabili. Volendo rilassare la condizione precedente in: "$0 < y_{i1} + 2 < y_{i2}$", per ammettere la non indennizzabilità dei giorni sabato e domenica applicata a talune forme contrattuali, persistono forti incoerenze in 169 osservazioni. In realtà, in dieci casi si tratta di incoerenze apparenti (errori di copertura delle dichiarazioni a conguaglio), poiché associate a dichiarazioni attestanti la differenza di accredito. Per le rimanenti 159 dichiarazioni si è deciso di stimare la sola componente economica legata al periodo di carenza, ammettendo la possibilità che la mancata erogazione della prestazione sia connessa ad accertamenti che hanno dato luogo al rigetto. Per altro verso, si riscontrano 129 violazioni legate a importi positivi a conguaglio che non trovano un evento corrispondente sull'archivio telematico della malattia (*missing value*). In tale circostanza si nota una concentrazione di casi nel mese di gennaio (57%), a indicare un possibile trascinamento in avanti degli importi anticipati riferibili all'anno precedente. Alla luce di queste considerazioni, si è stabilito di includere nella stima finale i valori degli importi dichiarati nell'archivio Emens, a meno di quelli presentati nel mese di gennaio (esclusi pure ai fini della carenza), per una numerosità finale pari a 56 dichiarazioni.

In definitiva, nella stima diretta dei trattamenti di malattia anticipati dai datori per conto dell'INPS concorrono 4.024 record (a cui andrebbero aggiunti 10 record da rielaborare con la fonte DIFFACR), mentre per la stima indiretta del periodo di carenza si debbono aggiungere ai precedenti altre 2.174 unità[12], per le quali l'indennizzo è esclusivo carico del datore. Nel procedimento indiretto è necessario recuperare l'informazione relativa alla retribuzione media giornaliera, la quale non sempre è rilevabile dalle fonti previdenziali. In questa circostanza l'archivio fiscale della Certificazione Unica costituisce una valida soluzione. Stesso discorso si applica ai dipendenti pubblici indennizzati che ammontano a 3.074 unità. Infine, per quanto concerne le prestazioni dirette di malattia a carico INPS si osserva la presenza di 194 titolari, il cui importo pagato non pone problemi computazionali.

La fase di validazione esterna prevede come primo punto il confronto tra la stima campionaria del numero di giorni di malattia dei dipendenti pubblici e il *benchmark* rappresentato dal macro-dato ricavato del Conto annuale della Ragioneria Generale dello Stato (RGS). I dipendenti pubblici del Conto annuale 2015 ammontavano a 3,26 milioni ed erano costituiti per il 43,8% da uomini e il restante 56,2% da donne. Per la stima finale di IT-SILC14_15 si sono utilizzati due approcci. Il primo applica ai dati osservati un fattore di riporto all'universo (N/n) distinto per sesso, utilizzando la popolazione degli assicurati nel

11 In realtà sussiste un terzo vincolo che impone, in aggiunta al punto [2], la compatibilità dei mesi di erogazione e di manifestazione dell'evento morboso, con relativa rimozione delle casistiche dove l'erogazione avviene nei primi tre mesi dell'anno con inizio gennaio (si veda Tavola 11.7).

12 Il valore 2.174 si ricava dalla somma di 1.807 (periodo di carenza capiente) e 367 (377 periodi teoricamente indennizzabile ma senza dichiarazioni a conguaglio da cui si sottraggono i 10 recuperi dalle differenze di accredito).

settore pubblico, come da fonte INPS_GDP[13]. Il secondo utilizza i pesi diretti del campione IT-SILC14_15 per calcolare sia la stima del numero di giorni di malattia che la numerosità dei dipendenti pubblici (denominatore nel calcolo della media). In particolare la percentuale di maschi nel pubblico impiego è leggermente più elevata nella totalità della fonte previdenziale (43,1%) rispetto al campione pesato (41,5%). I due metodi di stima generano risultati simili. Dal confronto finale non emergono differenze significative sul totale e sulla media dei giorni retribuiti tra i dati stimati e quello RGS, mentre vi sono differenze di rilievo nella distribuzione per genere (Tavola 11.8). Ai fini dell'elaborazione si tenuto conto del periodo di comporto e pertanto si è effettuato un *top-coding* sui periodi di malattia superiori ai 180 giorni.

Tavola 11.8 - Stime IT-SILC14_15 e fonte RGS: numero totale e medio giorni di malattia, per sesso. Anno 2015

Fonte	Totale giorni malattia retribuiti			Media(a) giorni retribuiti		
	M	F	MF	M	F	MF
RGS	13.496.092	18.676.120	32.172.212	9,50	10,23	9,9
GDP_ITS_14_15(b)	10.860.012	21.337.406	32.197.418	7,35	10,96	9,4
GDP_ITS_14_15W(c)	10.269.162	21.755.066	32.024.228	7,26	10,90	9,4

(a) intesa come rapporto tra numero totale di giorni di malattia e numero totale di lavoratori pubblici.
(b) la stima applica i fattori di riporto, distinti per sesso, ricavati dalla fonte INPS_GDP (3,4 mln dipendenti).
(c) la stima utilizza i pesi diretti del campione IT-SILC14_15.

La procedura diretta applicata alle prestazioni anticipate dai datori lavoro per conto dell'INPS determina il seguente risultato: a livello campionario la media degli importi è pari 568 euro, per una spesa totale di circa 2,3 milioni euro destinata a 4.024 unità campionarie (2° rigo di Tavola 11.9).

Tavola 11.9 - Stima degli importi di indennità malattia anticipata per conto dell'INPS Anno 2015

Fonti del dato	Numero osserv. dip. priv.	Coef. di riporto	Stima numero dipendenti privati	% distors. rispetto Uniem	Corr. distors	Nr. osserv. titolari malat.	Importi malattia (migl. €)	Stima importi malattia (migl. €)	Stima corretta importi malattia (migl. €)
UNIEMES_UNIV	14.885.466	-	-	-	-	-	-	-	-
UNIE_ITS14_15	22.033	6.755.987	14.885.466	0%	1	4.024	2.286	**1.544.104**	-
UNIE_ITS14_15W	21.050	Pesi diretti	13.998.240	5,96%	10.634	3.833	2.201	1.467.005	**1.559.985**

Per stimare la spesa relativa all'indennità di malattia anticipata, sono state adottate due strategie. Nella prima si è semplicemente utilizzato un fattore di riporto all'universo (N/n) ai dati osservati, prendendo a riferimento la popolazione dei lavoratori del privato che presentano almeno una dichiarazione mensile (con imponibile previdenziale positivo) nell'archivio Uniemens[14]. Nella seconda, ritenuta migliore, sono stati applicati i pesi diretti del campione IT-SILC14_15, corretti per la distorsione rispetto alla popolazione Uniemens dei dipendenti del privato (di cui al passo precedente). Come si evince dalla Tavola 11.9 i due criteri generano risultati molto simili in termini di spesa stimata, rispettivamente pari a: 1,54 e 1,56 miliardi di euro.

13 Tale sigla si riferisce alla banca dati INPS-Gestione Dipendenti Pubblici alimentata dai flussi informativi delle amministrazioni dello Stato e gli altri enti pubblici riguardanti i dati anagrafici e previdenziali dei propri dipendenti. I microdati della suddetta fonte sono stati incrociati con i dati campionari EU-SILC14-15 sia per il calcolo del fattore di riporto e sia per identificare il collettivo dei dipendenti pubblici a cui applicare i pesi diretti.

14 Il numero totale di lavoratori del privato, con imponibile previdenziale positivo, nell'universo Unimens ammonta a 14 milioni e 900 mila unità, mentre nel campione (intersezione di Unimens ed EU-SILC14-15) è pari a 22.033 unità. Dal rapporto tra queste due grandezze si ricava un coefficiente di riporto del valore di 675,9787.

Applicando la seconda strategia anche ai dati sui trattamenti di malattia a erogazione diretta INPS (Tavola 11.5), da rapportare però alla popolazione dei beneficiari di fonte INPS "prestazioni dirette non pensionistiche", si perviene a un valore stimato di 255 milioni di euro.

In conclusione, i dati sperimentali del 2015 stimano una spesa pari a circa 1,815 miliardi di euro per le prestazioni di malattia diretta e anticipata, a fronte di 1,880 miliardi in pagamenti per le stesse voci, desumibili dall'indagine sui bilanci consuntivi degli Enti previdenziali 2015 (*benchmark*)[15]. Volendo trarre un bilancio finale della sperimentazione, possiamo concludere che i risultati ottenuti si conformano ai valori dei rispettivi *benchmark*, avvalorando l'ipotesi di una sua implementazione a regime.

11.6. Conclusioni

Il documento si propone di fornire all'utente, anche non esperto in materia, una chiave di lettura semplice, non eccessivamente tecnica, della teoria che sta alla base del processo di elaborazione statistica dell'informazione, utilizzando come caso studio la prestazione di malattia nelle sue varie articolazioni. Prendendo spunto dalla sperimentazione della fonte sulle certificazioni telematiche della malattia, si è voluto enfatizzare lo schema logico che presiede il trattamento del dato statistico nell'ottica di una sua possibile diffusione. Tale schema parte dalla definizione concettuale del fenomeno, si sviluppa attraverso lo studio delle fonti (metadati), l'individuazione dei criteri più appropriati di calcolo e relative ipotesi di lavoro e, infine, termina con la verifica empirica della bontà dei vari metodi a confronto, supportata dalla validazione esterna dei risultati con le fonti ufficiali (*benchmark*). La sperimentazione dell'uso di nuove fonti del dato si realizza in un ciclo incrementale, dove la conoscenza del fenomeno basato sulle fonti esistenti si arricchisce sempre più di nuovi contenuti informativi, permettendo al tempo stesso di migliorare la qualità dei dati. Il disegno multi-fonte, presentato in questo lavoro, si innesta in questo modello incrementale, prevedendo l'affiancamento delle informazioni lungo una pluralità di archivi, con la relativa validazione dei dati combinati (*common variables validation*). Alla base di quest'ultima fase vi è, poi, l'implementazione di un sistema di vincoli relazionali che legano le variabili oggetto di studio attraverso le varie fonti, assicurando così la loro consistenza. Il disegno multi-fonte fornisce la soluzione metodologica ottimale quando ci si trova di fronte a un'informazione frammentata e diffusa su una moltitudine di archivi e resa ulteriormente complessa dalla varietà di regimi che sovrintendono il sistema e le modalità di erogazione.

Questa sperimentazione si colloca, oltre che sul filone della ricerca delle indagini sociali sui redditi (EU-SILC), anche su quello più generale dello sviluppo di statistiche sui trattamenti monetari non pensionistici, seppure con un taglio dedicato più specificatamente alle statistiche sociali. Nei contenuti, tuttavia, mantiene la sua validità anche in relazione a eventuali analisi di tipo economico, qualora si intendesse realizzare uno studio per settore ATECO o classe dimensionale d'impresa. La qual cosa però implica una diversa selezione del campione (o la rilevazione totale); ad esempio, tutti i lavoratori dipendenti appartenenti a determinati gruppi di imprese[16]. Il vantaggio che offre la presente sperimentazione, rispet-

15 Occorre precisare che il termine di paragone qui utilizzato (*benchmark*) dovrebbe sovrastimare, seppur di poco, la consistenza del fenomeno; in quanto nelle voci di bilancio dell'Ente sono inclusi anche gli arretrati. Nella sperimentazione, invece, si è adottato un criterio di competenza riferito al momento in cui si manifesta l'evento morboso, da cui l'eliminazione dei importi conguagliati riferibili all'anno precedente.

16 In proposito si può citare il lavoro di Consolini, 2004 che sfrutta il contenuto dei dati 770 Agenzia delle Entrate per la stima

to ad altre tecniche utilizzate in passato, è quella di fornire la massima disaggregazione del dato lungo i vari livelli di classificazione (lavoratore e relative caratteristiche socio-economiche, tipologia di impresa o istituzione pubblica presso cui lavora, tipo di trattamento o di regime di che eroga la prestazione, eccetera).

L'applicazione pratica della nuova tecnica di stima dell'indennità di malattia, calata nell'attuale sistema italiano di calcolo delle componenti di reddito IT-SILC, produrrà inevitabilmente una rottura rispetto al passato e richiederà la revisione all'indietro della serie storica sui dati d'indagine. Il beneficio futuro atteso è di mettere a disposizione dell'utente finale la rappresentazione di un fenomeno altrimenti nascosto e, non ultimo, di migliorare la riconciliazione micro-macro dei dati reddituali che attualmente sta investendo il settore delle indagini sociali e della contabilità nazionale.

dei Trattamenti Monetari Non Pensionistici (TMNP), in un'ottica orientata sul dato economico. Si può prendere, inoltre, a riferimento il lavoro di Spinelli e Tancioni M., 2004, che affronta, sempre con un taglio orientato all'analisi economica, la problematica del trattamento degli archivi amministrativi DM10 dell'INPS (relativi al triennio 1999-2001), presi nella loro totalità, per il calcolo degli importi a livello d'impresa sui TMNP anticipati dal datore di lavoro per conto dell'INPS.

12. L'UTILIZZO DELLA BANCA DATI REDDITUALE DEL MEF PER LA CALIBRAZIONE DEL CAMPIONE[1]

12.1 Introduzione

Il tentativo di questo lavoro è quello di valutare la presenza di distorsioni nelle stime dell'indagine EU-SILC relative ai redditi e ai principali indicatori distributivi (rischio di povertà e indicatori di disuguaglianza), connesse ad un'imperfetta rappresentazione della "vera" struttura reddituale della popolazione.

L'ipotesi è che il meccanismo della mancata risposta totale delle unità di rilevazione, le famiglie, non sia del tutto indipendente dalle variabili oggetto di analisi e cioè dai redditi, nemmeno controllando e correggendo gli effetti sulla mancata risposta tramite le informazioni anagrafiche, disponibili sia per le famiglie rispondenti, sia per quelle non rispondenti.

L'idea è di fare in modo che il campione EU-SILC riesca a riprodurre correttamente le informazioni provenienti, in maniera del tutto esogena, dalla Banca Dati Reddituale (BDR), messa a disposizione dell'Istat dal MEF a partire dal 2011. In tale registro sono presenti le principali informazioni provenienti dalle dichiarazioni fiscali sulle persone fisiche, sintetizzate a livello di singolo individuo (oltre 40 milioni di record per anno).

È importante sottolineare che il reddito fiscale presente nella BDR non corrisponde al reddito disponibile (utilizzato dalle famiglie per il consumo o per il risparmio) oggetto principale dell'indagine EU-SILC. Questo per una molteplicità di ragioni, delle quali possiamo segnalare le più rilevanti:

- Presenza di fasce di esenzione, tali che non tutte le componenti del reddito disponibile debbano essere dichiarate al Fisco;
- Esistenza di una quota di redditi "sommersi" che sfuggono alla rilevazione fiscale ma che si ritiene di poter rilevare, almeno parzialmente, tramite l'intervista, confidando nell'efficacia delle rassicurazioni fornite ai rispondenti circa l'assoluta riservatezza garantita dall'Istat;
- Presenza di tipologie reddituali (e.g. alcuni redditi da capitale) che sono tassati alla fonte e che non sono presenti nelle dichiarazioni fiscali e quindi nella BDR;
- Presenza di componenti reddituali esenti dalla tassazione (ad esempio assegni familiari, pensioni invalidità civile, eccetera) e della stessa imposta (essendo il reddito fiscale assimilabile al concetto di reddito imponibile).

Tuttavia, ipotizzando che il reddito dichiarato al fisco presente in BDR sia positivamente correlato con il reddito disponibile ottenuto con l'indagine, un processo di calibrazione dei pesi di EU-SILC basato anche sui redditi fiscali, tale, cioè, di riprodurre i totali noti di BDR, permetterebbe di contenere la distorsione finale delle stime sui redditi disponibili ricavate dalla stessa indagine.

L'operazione che si vuole effettuare, quindi, è di associare ad ogni record del campione teorico (inclusivo anche degli individui appartenenti a famiglie non rispondenti) le informa-

1 I paragrafi 12.1, 12.2 e 12.3 sono stati redatti da Silvano Vitaletti, mentre i paragrafi 12.4, 12.5, 12.6 sono stati redatti da Stefano Gerosa. Le conclusioni 12.7 sono state redatte in comune dai due curatori.

zioni dichiarate al fisco e presenti nella BDR. In questo modo sarà possibile: a) verificare se e in che misura il campione finale dell'indagine sia in grado, tramite i coefficienti di riporto all'universo già calcolati nell'indagine a regime, di riprodurre i totali di BDR; b) tentare di correggere le eventuali distorsioni, sia nella fase di correzione dei pesi campionari in funzione della probabilità di risposta, includendo alcune variabili di provenienza fiscale nei modelli di stima, sia nella fase finale di calibrazione dei pesi, vincolando i pesi campionari finali a riprodurre alcuni totali noti della BDR.

12.2 La scelta delle variabili della BDR

Le variabili considerate sono, oltre al reddito complessivo, il reddito da lavoro dipendente, il reddito da pensione e quello da lavoro autonomo: insieme tali redditi rappresentano circa il 96% del reddito fiscale totale. Tutte le componenti di reddito sono state considerate al lordo delle tasse e prima di eventuali deduzioni. Per il reddito da lavoro autonomo non sono state considerate le componenti negative, ovvero le eventuali perdite, in quanto ritenute occasionali e non rappresentative del reale livello reddituale degli individui. Per lo stesso motivo, le componenti negative sono state stralciate anche dal reddito complessivo. La tavola 12.1 fornisce il dettaglio delle voci reddituali incluse nell'analisi.

Tavola 12.1 - Le variabili BDR incluse nell'analisi

NOME	REDDITI	NOME BDR	2011	2012	2013	2014	2015
DIP_BDR	Redditi di lavoro dipendente	RC005DIP	X	X	X	X	X
DIP_BDR	Redditi di lavoro dipendente dei frontalieri - Quota esente	RC005001		X	X	X	X
DIP_BDR	Premi di produzione assoggettati ad imposta sostitutiva	PREMI_IMP_SOST	X	X	X	X	
DIP_BDR	Altri redditi assimilati a quelli di lavoro dipendente	RC009001F0	X				
		RC009002F0		X	X	X	X
PEN_BDR	Redditi di pensione	RC005PEN	X	X	X	X	X
AUT_BDR	Redditi dell'impresa agricola o di allevamento di spettanza dell'imprenditore al netto delle perdite d'impresa in contabilità ordinaria	RD018001	X	X	X	X	X
AUT_BDR	Redditi di lavoro autonomo derivanti dall'esercizio di arti e professioni	RE025001P	X	X	X	X	X
AUT_BDR	Redditi d'impresa di spettanza dell'imprenditore al netto delle perdite d'impresa in contabilità ordinaria	RF051001	X	X			
		RF101001			X	X	X
AUT_BDR	Redditi d'impresa di spettanza dell'imprenditore al netto delle perdite d'impresa in contabilità semplificata	RG034001P	X	X			
		RG036001P			X	X	X
AUT_BDR	Totale redditi di partecipazione in società di persone e assimilate esercenti attività d'impresa	RH014002P	X	X	X	X	X
AUT_BDR	Totale redditi di partecipazione in associazioni tra artisti e professionisti	RH017001P	X	X	X	X	X
AUT_BDR	Totali redditi di partecipazione in società semplici	RH018001	X	X	X	X	
		RH018001P					X
AUT_BDR	Redditi derivanti da attività occasionale o da obblighi di fare, non fare e permettere	RL019001	X	X	X	X	X
AUT_BDR	Redditi derivanti da attività sportive dilettantistiche e collaborazioni con cori, bande e filodrammatiche	RL022002	X	X	X	X	X
AUT_BDR	Altri redditi di lavoro autonomo	RL030001	X	X	X	X	X
AUT_BDR	Redditi di lavoro autonomo da mod. 770	RDT_AU	X	X	X	X	X
TOT_BDR	Reddito complessivo (*)	RN001005	X	X	X	X	X

(*) Dal reddito complessivo sono state escluse le componenti negative del reddito da lavoro autonomo (RE025001N, RG034001N/RG036001N, RH014002N, RH017001N). In questo modo il reddito complessivo risulta sempre positivo.

12.3 Inserimento delle variabili della BDR nel campione EU-SILC

La possibilità di collegare i record individuali del campione EU-SILC ai record individuali della BDR, dalla quale prelevare le informazioni sui redditi fiscali, viene fornita dalla disponibilità dei sistemi di integrazione delle fonti amministrative (SIM - il Sistema Integrato dei Microdati di fonte amministrativa) e di indagine (SIRIL – Sistema Integrato delle Rilevazioni su individui e famiglie) che l'Istat ha realizzato nel corso degli anni.

Il sistema SIM è stato costruito dall'Istat per rendere disponibili e utilizzabili, all'interno dell'Istituto per la produzione delle statistiche ufficiali e in forma del tutto anonima, i microdati di fonte amministrativa. In sostanza, a ciascun individuo presente negli archivi amministrativi integrati nel sistema, viene assegnata una chiave anonima e stabile nel tempo (il codice individuo SIM) che lo identifica univocamente nei diversi archivi integrati. In questo modo diventa possibile utilizzare congiuntamente le informazioni provenienti da archivi differenti che afferiscono ad uno stesso individuo. Il sistema SIM è stato costruito e viene costantemente aggiornato in modo incrementale (le prime fonti integrate risalgono al 2009) attraverso procedure di record linkage che, quando il soggetto presente in una nuova fonte viene identificato tra quelli già presenti nel sistema, gli assegnano lo stesso codice. Ad ogni nuovo individuo (non identificato tra quelli già presenti) viene invece assegnato un nuovo codice, disponibile per integrazioni successive di altre fonti.

In tempi più recenti (la prima rilevazione integrata, nel 2017, è stata proprio l'indagine EU-SILC), ma in modo analogo, è stato dato avvio alla costruzione del sistema SIRIL. In questo nuovo sistema, gli individui appartenenti ai campioni delle indagini che, attraverso procedure di *record linkage*, vengono identificati all'interno del sistema SIM, ricevono lo stesso codice già presente in SIM. Per le indagini, a differenza degli archivi amministrativi presenti in SIM, non vengono attribuiti dei codici nuovi agli individui non identificati[2].

Agli individui del campione privi di un codice SIM (non identificati), questo è stato imputato con il metodo del donatore più simile[3] appartenente allo stesso campione. Il numero di casi in cui il codice_individuo è stato imputato è mostrato nella Tavola 12.2.

In questo modo, i soggetti rilevati con l'indagine che non sono stati identificati possono ricevere da BDR le informazioni sui redditi fiscali proprie dei soggetti a loro più simili (eventualmente nessuna, quando il donatore non è un soggetto dichiarante e quindi non presente in BDR).

2 Si ritiene che una tale operazione possa considerarsi inutile per gli individui appartenenti ai campioni teorici, estratti da fonti amministrative già integrate in SIM (in genere si tratta delle LAC - Liste Anagrafiche comunali) e quindi già presenti in SIM. Tale operazione può invece essere considerata pericolosa per gli individui rilevati che non vengono riconosciuti tra quelli delle rispettive famiglie anagrafiche, in quanto provvisti di dati identificativi piuttosto poveri (a seconda dell'indagine) e di dubbia qualità (rilevati nel corso delle interviste), che se inseriti nel sistema potrebbero dare luogo a una propagazione di possibili errori di linkage, con l'integrazione di nuove fonti che entrano nel sistema.

3 Il processo di donazione è stato realizzato mediante la procedura FRI (Full Record Imputation) sviluppata in Istat da Pierpaolo Massoli e basata su una valutazione della similarità tra due unità statistiche utilizzando l'indice di Gower, il quale può essere adattato per l'utilizzo congiunto di variabili sia qualitative che quantitative. La ricerca del donatore di distanza minima è stata effettuata all'interno di strati definiti dalle seguenti classificazioni: regione; tipo di comune (centro area metropolitana, periferia area metropolitana, oltre 50.000 ab., 10.001-50.000 ab., 2.001-10.000 ab., fino a 2.000 ab.); classe di età (0-15, 16-24, 25-34, 35-44, 45-64, 65 e oltre); sesso; cittadinanza dell'intestatario della scheda di famiglia anagrafica (italiana, straniera EU, straniera non EU) ; paese di nascita (Italia, estero EU, estero non EU). Nel primo di due step, l'unica variabile considerata per misurare la similarità tra i soggetti (donatore-ricevente) è stata il numero componenti della famiglia (1, 2, 3, 4, 5 o più). Nel secondo step sono stati trattati i casi residui, includendo nel calcolo della similarità anche la cittadinanza dell'ISF e il paese di nascita, ovviamente rimosse dalla definizione degli strati.

Tavola 12.2 - Imputazione da donatore del codice_individuo SIM nel campione EU-SILC

	ANNO DI INDAGINE				
	2012	2013	2014	2015	2016
Numero individui campione (*)	64.162	64.633	64.679	63.662	68.810
Numero individui con codice_individuo SIM imputato	411	295	252	229	257
Percentuale di imputazione	0,64	0,46	0,39	0,36	0,37

(*) Sono compresi sia gli individui del campione teorico, cioè tutti i componenti della famiglia anagrafica estratta, sia gli eventuali altri componenti rilevati nella famiglia di fatto.

12.4 Confronto delle stime EU-SILC con i totali noti della BDR

Aver riportato direttamente nel campione EU-SILC le stesse variabili di reddito dichiarate all'Agenzia delle Entrate ci permette di valutare la capacità del campione stesso di riprodurre gli aggregati dell'universo BDR: se non vi fossero distorsioni significative e se il campione rappresentasse accuratamente la "popolazione fiscale", allora le stime campionarie degli aggregati fiscali non dovrebbero differire in modo sistematico dai totali BDR.

Utilizzando i coefficienti di riporto all'universo calcolati nell'indagine EU-SILC corrente si osserva (Tavole 12.3, 12.4 e 12.5) una distorsione piuttosto accentuata delle stime per quasi tutte le categorie di reddito considerate.

Si può anche osservare che la distorsione si modifica nettamente in ampiezza in coincidenza della riprogettazione dell'indagine occorsa nel 2016 (utilizzo di una tecnica mista CAPI-CATI[4], differente modulazione del questionario, nuova Società affidataria della rilevazione, nuova rete di rilevatori).

L'ammontare di reddito fiscale complessivo risulta sottostimato lungo tutto il periodo considerato, per una frazione compresa tra il 2,5% e il 4,4%, mentre il numero di dichiaranti di una qualunque tipologia di reddito presenta in genere una sottostima di entità lievemente superiore: ne segue una stima campionaria del reddito fiscale medio per singolo dichiarante superiore al dato di origine amministrativa, tranne negli anni 2014 e 2015.

Tavola 12.3 - Confronto degli ammontari totali di reddito fiscale per tipologia, da BDR e da stime campionarie EU-SILC. Totale nazionale (MLN di euro)

		2012	2013	2014	2015	2016
Reddito totale	Universo BDR	802.987	797.320	806.091	810.642	824.046
	Stima EU-SILC	771.819	772.777	773.889	775.104	803.748
	Differenza %	-3,88	-3,08	-3,99	-4,38	-2,46
Reddito da lavoro dipendente	Universo BDR	432.130	429.907	427.193	426.271	433.825
	Stima EU-SILC	426.055	430.993	427.099	424.066	435.294
	Differenza %	-1,41	0,25	-0,02	-0,52	0,34
Reddito da lavoro autonomo	Universo BDR	106.142	107.713	104.957	105.198	106.302
	Stima EU-SILC	91.476	92.600	87.273	89.249	96.938
	Differenza %	-13,82	-14,03	-16,85	-15,16	-8,81
Reddito da pensione	Universo BDR	233.864	238.810	243.617	247.201	249.218
	Stima EU-SILC	226.362	231.344	233.643	233.970	238.361
	Differenza %	-3,21	-3,13	-4,09	-5,35	-4,36

4 Anche nella rilevazione 2015 è stata parzialmente utilizzata la tecnica CATI ma in una quota molto più modesta di famiglie (cfr. capitolo 1.2 per dettagli sulla transizione a tecnica mista).

Tavola 12.4 - Confronto del numero di dichiaranti per tipologia di reddito, da BDR e da stime campionarie EU-SILC. Totale nazionale (migliaia)

		2012	2013	2014	2015	2016
	Universo BDR	40.833	40.351	40.242	40.067	40.101
Reddito totale	Stima EU-SILC	38.983	38.646	38.679	38.529	38.275
	Differenza %	-4,53	-4,23	-3,88	-3,84	-4,55
	Universo BDR	21.084	20.914	20.573	20.580	20.990
Reddito da lavoro dipendente	Stima EU-SILC	20.472	20.457	20.231	20.324	20.552
	Differenza %	-2,90	-2,19	-1,66	-1,24	-2,08
	Universo BDR	5.933	6.320	6.193	6.179	5.961
Reddito da lavoro autonomo	Stima EU-SILC	5.418	5.863	5.700	5.709	5.667
	Differenza %	-8,68	-7,23	-7,96	-7,61	-4,94
	Universo BDR	15.064	15.131	14.963	14.799	14.774
Reddito da pensione	Stima EU-SILC	14.281	14.253	14.150	13.736	13.600
	Differenza %	-5,20	-5,80	-5,43	-7,18	-7,95

Tavola 12.5 - Confronto dei redditi fiscali medi per tipologia, da BDR e da stime campionarie EU-SILC. Totale nazionale (Euro)

		2012	2013	2014	2015	2016
	Universo BDR	19.665	19.759	20.031	20.232	20.549
Reddito totale	Stima EU-SILC	19.799	19.996	20.008	20.117	20.999
	Differenza %	0,68	1,20	-0,11	-0,57	2,19
	Universo BDR	20.495	20.556	20.765	20.713	20.669
Reddito da lavoro dipendente	Stima EU-SILC	20.812	21.068	21.111	20.865	21.180
	Differenza %	1,54	2,49	1,67	0,74	2,48
	Universo BDR	17.891	17.042	16.947	17.026	17.832
Reddito da lavoro autonomo	Stima EU-SILC	16.884	15.794	15.311	15.633	17.106
	Differenza %	-5,63	-7,32	-9,66	-8,18	-4,07
	Universo BDR	15.524	15.783	16.281	16.704	16.869
Reddito da pensione	Stima EU-SILC	15.851	16.231	16.512	17.033	17.527
	Differenza %	2,10	2,84	1,42	1,97	3,90

La disaggregazione del reddito fiscale complessivo nelle sue principali componenti mostra ampie differenze nelle stime campionarie delle diverse tipologie reddituali e consente di valutare il contributo relativo del numero di dichiaranti e dei redditi medi alla distorsione totale.

Il reddito fiscale complessivo da lavoro dipendente è stimato con relativa precisione, e la differenza massima con il totale di BDR è pari all'1,4%: il totale dei dichiaranti è però sottostimato lungo tutto il periodo mentre i redditi fiscali medi sono sempre maggiori nel campione rispetto alla popolazione BDR.

Una generale e molto marcata sottostima affligge poi i redditi da lavoro autonomo: il totale di fonte campionaria è inferiore al totale BDR per percentuali che vanno dal 9% circa del 2016 al 17% del 2014. Ciò è dovuto sia ad una sottostima del numero dei dichiaranti (che varia dal 5% al 9% circa) che a redditi fiscali medi campionari più bassi di quelli osservati nella BDR.

Anche per i redditi da pensione osserviamo una sottostima che varia dal 3% al 5% circa del totale, che in questo caso dipende da una significativa sottostima del numero di dichiaranti (che varia dal 5% all'8% circa), mentre i redditi fiscali medi stimati da campione sono più alti di quelli calcolati sul totale della popolazione BDR.

La Tavola 12.6 mostra una decomposizione delle differenze tra BDR e EU-SILC nel numero complessivo di dichiaranti reddito fiscale secondo tre gruppi di percettori: per ogni tipologia di reddito considerata, i dichiaranti vengono distinti in un gruppo a basso reddito (1° Quartile), un gruppo centrale (2° e 3° Quartile) e un gruppo ad alto reddito (4° Quartile). È così possibile notare come le distorsioni non siano distribuite in modo omogeneo, ma siano invece concentrate in particolari sotto-gruppi.

Relativamente al reddito fiscale totale, è il gruppo dei contribuenti a basso reddito ad essere il più sotto-rappresentato, per una percentuale variabile tra il 12% e il 15%. I contribuenti del gruppo centrale sono ben rappresentati, con uno scarto massimo dell'1,6%, mentre i percettori di redditi elevati sono in generale sottorappresentati per una frazione che arriva fino al 3%. Vi sono però importanti differenze anche tra le diverse tipologie di reddito fiscale considerate. Per i redditi fiscali da lavoro dipendente, i contribuenti del gruppo a basso reddito sono gli unici ad essere sottorappresentati (fino a un massimo del 15,4% nel 2012), mentre gli altri due gruppi di contribuenti sono al contrario sovra-rappresentati nel campione EU-SILC: ad esempio il gruppo dei contribuenti ad alto reddito è stimato essere del 5% più numeroso rispetto al totale BDR nel 2016. Per i redditi da lavoro autonomo, al contrario, è il gruppo dei contribuenti ad alto reddito ad essere il più sottorappresentato nell'indagine, per una percentuale che arriva fino al 15,2% nel 2015, mentre i contribuenti degli altri due gruppi sono sottorappresentati in modo variabile lungo il periodo considerato per valori tra il 3% e il 9% circa. I percettori di redditi da pensione sono sotto-rappresentati per tutti i gruppi e lungo tutto il periodo, in particolare quelli del primo quartile con punte del 18% circa nel 2016.

L'ampiezza delle distorsioni campionarie qui evidenziate mostra quanto sia importante definire una strategia per il loro contenimento, anche in relazione alla dimostrata variabilità della distorsione stessa rispetto a cambiamenti nella conduzione della rilevazione e dunque alla confrontabilità delle stime nel tempo. Una strategia di riduzione delle distorsioni dovrà anche tenere conto della loro complessa articolazione in funzione delle tipologie di reddito e del livello del reddito del contribuente.

Ottenere stime corrette sui redditi fiscali dal campione di indagine non è, tuttavia, di per sé sufficiente. Occorre che la correzione della distorsione sui redditi fiscali porti con sé anche una riduzione della distorsione delle stime sui redditi disponibili e degli indicatori associati alla distribuzione di quest'ultima variabile.

Tale distorsione non è effettivamente nemmeno misurabile, dal momento che non esistono *benchmark* utili allo scopo (se esistessero potrebbero essere direttamente utilizzati per la correzione della distorsione o addirittura sostituire la rilevazione stessa).

Possiamo tuttavia affermare che agire sulla correttezza delle stime sui redditi fiscali migliori, sebbene in modo indiretto, la correttezza delle stime sui redditi disponibili, pur essendo difficile quantificare il beneficio. Questo perché tra il reddito fiscale e il reddito disponibile oggetto di indagine esiste una forte correlazione positiva, nonostante siano costruite in maniera diversa.

Tavola 12.6 - Confronto del numero di dichiaranti di reddito fiscale per quartile di reddito e tipologia, da BDR e da stime campionarie EU-SILC. Totale nazionale (migliaia di individui)

		2012	2013	2014	2015	2016
		DICHIARANTI DI REDDITO				
	BDR	10.208	10.088	10.059	10.015	10.023
1° QUARTILE	EU-SILC	8.732	8.636	8.844	8.824	8.504
	Differenza %	-14,46	-14,39	-12,08	-11,89	-15,15
	BDR	20.417	20.176	20.122	20.035	20.053
2° E 3° QUARTILE	EU-SILC	20.341	20.025	20.023	19.988	19.724
	Differenza %	-0,37	-0,75	-0,49	-0,23	-1,64
	BDR	10.208	10.088	10.061	10.017	10.026
4° QUARTILE	EU-SILC	9.910	9.985	9.812	9.717	10.047
	Differenza %	-2,92	-1,02	-2,48	-3,00	0,21
		DICHIARANTI DI REDDITO DA LAVORO DIPENDENTE				
	BDR	5.272	5.229	5.143	5.146	5.247
1° QUARTILE	EU-SILC	4.458	4.464	4.371	4.516	4.541
	Differenza %	-15,43	-14,63	-15,02	-12,23	-13,46
	BDR	10.542	10.457	10.287	10.290	10.495
2° E 3° QUARTILE	EU-SILC	10.691	10.666	10.654	10.624	10.493
	Differenza %	1,41	2,00	3,57	3,25	-0,02
	BDR	5.271	5.229	5.143	5.145	5.247
4° QUARTILE	EU-SILC	5.323	5.327	5.206	5.184	5.518
	Differenza %	0,99	1,88	1,23	0,76	5,16
		DICHIARANTI DI REDDITO DA LAVORO AUTONOMO				
	BDR	1.483	1.590	1.547	1.543	1.489
1° QUARTILE	EU-SILC	1.433	1.563	1.405	1.416	1.370
	Differenza %	-3,37	-1,69	-9,19	-8,25	-7,98
	BDR	2.966	3.150	3.097	3.090	2.981
2° E 3° QUARTILE	EU-SILC	2.678	2.905	2.977	2.983	2.927
	Differenza %	-9,72	-7,78	-3,88	-3,46	-1,83
	BDR	1.483	1.580	1.549	1.545	1.491
4° QUARTILE	EU-SILC	1.307	1.395	1.318	1.310	1.370
	Differenza %	-11,88	-11,73	-14,90	-15,22	-8,12
		DICHIARANTI DI REDDITO DA PENSIONE				
	BDR	3.766	3.783	3.741	3.700	3.694
1° QUARTILE	EU-SILC	3.195	3.144	3.262	3.098	3.037
	Differenza %	-15,17	-16,89	-12,80	-16,27	-17,78
	BDR	7.532	7.566	7.482	7.400	7.387
2° E 3° QUARTILE	EU-SILC	7.421	7.393	7.308	7.147	6.940
	Differenza %	-1,48	-2,28	-2,32	-3,41	-6,05
	BDR	3.766	3.782	3.741	3.700	3.693
4° QUARTILE	EU-SILC	3.665	3.716	3.580	3.491	3.623
	Differenza %	-2,68	-1,76	-4,30	-5,64	-1,90

In effetti, l'analisi della correlazione tra le variabili di reddito provenienti dalla BDR e quelle omologhe ottenute dall'indagine mostra che l'associazione è decisamente marcata e di segno positivo, soprattutto se si considera la correlazione per ranghi (Tavola 12.7). In sostanza, sebbene non si riscontri sempre una marcata relazione lineare (nel 2014 e nel 2015 la correlazione lineare tra i due redditi da lavoro autonomo flette a circa il 60%), quella per ranghi non scende mai al di sotto di circa il 75%. Ciò significa, in particolare, che le due distribuzioni sono conformi rispetto all'ordinamento: chi occupa una determinata posizione nella scala ordinata del reddito fiscale si trova in una posizione corrispondente (o vicina) nella scala ordinata del reddito fiscale.

Tavola 12.7 - Correlazione tra i redditi EU-SILC e i redditi BDR

Tipologia di reddito	Anno di indagine EU-SILC (*)				
	2012	2013	2014	2015	2016
	Coefficiente di correlazione lineare (ρ di Pearson)				
Reddito totale	0,76	0,76	0,82	0,77	0,76
Reddito da lavoro dipendente	0,8	0,59	0,61	0,88	0,8
Reddito da lavoro autonomo	0,85	0,83	0,88	0,84	0,85
Reddito da pensione	0,75	0,74	0,77	0,85	0,75
	Coefficiente di correlazione per ranghi (ρ_s di Spearman)				
Reddito totale	0,87	0,87	0,89	0,88	0,87
Reddito da lavoro dipendente	0,75	0,78	0,78	0,77	0,75
Reddito da lavoro autonomo	0,89	0,89	0,91	0,88	0,89
Reddito da pensione	0,87	0,87	0,87	0,86	0,87

(*) I redditi si riferiscono all'anno solare precedente quello di indagine.

Più in dettaglio, possiamo osservare che, a seconda delle diverse tipologie di reddito, la correlazione si posiziona su livelli differenti: quasi esatta quella relativa al reddito da lavoro dipendente, su un gradino più in basso quella relativa al reddito da pensione, ancora inferiore quella relativa ai redditi da lavoro autonomo.

Ciò che va sottolineato, infine, è che lo schema di correlazione (soprattutto quello per ranghi) rimane stabile negli anni considerati. Anche nel 2016, quando abbiamo visto mutare profondamente lo schema della distorsione, il legame tra reddito fiscale e reddito disponibile è rimasto sostanzialmente immutato.

12.5 La procedura di costruzione dei coefficienti di riporto all'universo e le variabili BDR

La procedura di costruzione dei coefficienti di riporto all'universo attualmente implementata per la componente trasversale dell'indagine EU-SILC è basata sull'uso di una famiglia di stimatori noti in letteratura come "stimatori di ponderazione vincolata" e consente la determinazione di un unico coefficiente di riporto all'universo per ciascuna famiglia e per ogni suo membro (calibrazione integrata), tali da produrre stime coerenti a un insieme di totali noti, desunti da fonti esterne.

La procedura inizialmente attribuisce ad ogni famiglia il peso diretto, dato dal reciproco della probabilità di inclusione come definita dal disegno di campionamento. Il peso diretto viene quindi modificato in due fasi distinte: nella prima si procede alla correzione del peso per la mancata risposta totale, nella seconda i pesi familiari così corretti vengono calibrati rispetto ad una serie di totali noti di fonte anagrafica, in modo da assicurarne la riproduzione da parte delle stime campionarie.

Nell'esercizio di seguito illustrato le variabili di fonte fiscale precedentemente descritte sono state introdotte in entrambe le fasi del processo di correzione del peso diretto iniziale.

12.5.1 Le variabili BDR nella correzione per la mancata risposta totale

Il peso diretto viene prima di tutto corretto per la probabilità di risposta, utilizzando le informazioni disponibili a livello familiare (per la prima *wave*) o individuale (per le *wave* successive alla prima). Infatti è difficile suppore che il meccanismo che genera la mancata risposta nell'indagine sia ignorabile, ovvero del tutto casuale: la variabilità dei tassi di rispo-

sta in relazione alle caratteristiche socio-demografiche e territoriali delle famiglie-campione sembra indicarlo, come molti studi condotti su indagini campionarie simili a EU-SILC[5].

Per le famiglie di prima *wave*, la probabilità di risposta viene stimata attraverso un modello di regressione logistica che usa esclusivamente informazioni ottenute dai registri anagrafici e include le variabili selezionate con un metodo di tipo "*stepwise*" da un insieme comprendente: tipo di dominio territoriale al livello NUTS II, ampiezza demografica del comune di residenza, numero dei componenti e caratteristiche dell'intestatario della scheda generale (sesso, età e cittadinanza). Per le *wave* successive alla prima, il peso iniziale (dato dal peso finale della precedente occorrenza di indagine) è corretto per la probabilità di risposta calcolata a livello individuale per tutti gli individui maggiori di 14 anni, e il modello di regressione logistica comprende questa volta anche covariate individuali (variabili indicatrici della presenza delle diverse tipologie reddituali, quintile di reddito di appartenenza, condizione professionale) e numerose informazioni a livello familiare rilevate nella precedente wave (tipologia familiare, titolo di godimento dell'abitazione, condizione di deprivazione materiale, rischio di povertà o bassa intensità lavorativa). La media dei pesi individuali corretti così ottenuti fornisce quindi il peso familiare corretto per la mancata risposta. In entrambi i casi per tenere conto della possibilità di diversi *pattern* di mancata risposta associate alle due tecniche di rilevazione utilizzate, vengono stimati modelli separati per i due sotto-campioni CAPI e CATI.

Le informazioni provenienti dalla BDR sono state quindi introdotte nella stima della probabilità di risposta per le famiglie della prima *wave*: la caduta delle famiglie si concentra infatti in prevalenza nella prima occasione di indagine, quando le uniche informazioni disponibili sono quelle provenienti dai registri anagrafici, mentre nelle occorrenze successive è già disponibile un ampio insieme di variabili di indagine per la correzione della mancata risposta[6].

In particolare per ogni famiglia è stato calcolato il reddito medio per percettore per ogni tipologia di reddito fiscale considerata e ogni famiglia è stata così classificata come a reddito basso, medio o alto secondo il quartile di appartenenza, e le variabili categoriali così calcolate sono state inserite nel modello di regressione logistica sopra descritto.

Tavola 12.8 - Stima degli odds ratio relativi alle variabili BDR nel modello logistico sulla probabilità di risposta delle famiglie campione (wave 1)

	2012		2013		2014		2015		2016			
	CAPI		CAPI		CAPI		CAPI/CATI		CAPI		CATI	
Dip. Q1	1,03		1		0,94		1,02		-		0,92	
Dip. Q2Q3	1,28	***	1,19	***	1,21	***	1,14	***	-		1,55	***
Dip. Q4	1,4	***	1,37	***	1,45	***	1,32	***	-		1,49	***
Pen. Q1	1,42	***	1,29	***	1,06		1,23	***	1,49	***	1,08	
Pen. Q2Q3	1,41	***	1,34	***	1,33	***	1,42	***	1,48	***	1,2	
Pen. Q4	1,66	***	1,61	***	1,71	***	1,61	***	1,33	***	1,62	***
Aut. Q1	1,29	***	1,29	***	-		-		0,79	***	-	
Aut. Q2Q3	1,13		1,07		-		-		1,08		-	
Aut. Q4	1,18		1,27	***	-		-		1,2	**	-	
N. osservazioni	10.080		10.027		10.043		9.981		9.674		4.311	

La categoria di riferimento è costituita dalle famiglie in cui non sono presenti percettori della tipologia di reddito considerata. (***, **): significativo al 99% e al 95%. (–): non selezionato dal metodo "stepwise".

5 Ad esempio Korinek, Mistiaen, and Ravallion (2006) per il Current Population Survey statunitense.

6 Tentativi di estendere l'uso delle variabili BDR alla correzione della mancata risposta nelle waves successive alla prima hanno in ogni caso fornito risultati molto simili a quelli qui descritti.

La tavola 12.8 mostra l'effetto di queste variabili sulla probabilità di risposta, come misurato dagli *odds ratio* del modello logistico: poiché per ogni tipologia reddituale, la categoria di riferimento scelta è quella di "famiglia non percettrice", un valore superiore a 1 segnala un'aumentata probabilità di risposta associata al fatto di essere una famiglia con reddito fiscale basso, medio o lato. Si può così osservare che le famiglie con redditi da lavoro dipendente o da pensione medio-alti hanno maggiore probabilità di partecipare all'indagine di quelle non percettrici o a reddito basso. Per i redditi da lavoro autonomo vi è una situazione più complessa: la significatività dell'associazione tra questi redditi fiscali e mancata risposta è in generale minore (nel 2015, nel 2016 e nel 2017 per la componente CATI tale variabile non è selezionata dal modello *stepwise*), e non si evidenzia lo stesso netto profilo crescente tra livello del reddito e probabilità di partecipazione all'indagine.

12.5.2 Le variabili BDR nella calibrazione del campione

I pesi corretti per la mancata risposta sono infine modificati attraverso due passi di calibrazione, che assicurano la riproduzione di un insieme di totali noti da parte delle stime campionarie, usando degli stimatori di ponderazione vincolata implementati attraverso il software Genesees, sviluppato internamente all'Istituto[7].

Nel primo passo il campione è calibrato in modo da riprodurre gli stessi totali stimati dalla Rilevazione campionaria sulle Forze di Lavoro circa la distribuzione per condizione occupazionale e titolo di studio della popolazione. Nel secondo sono invece usati una serie di totali noti provenienti dai registri demografici relativi alla distribuzione della popolazione per sesso, classe di età, cittadinanza, ampiezza demografica dei comuni a diversi livelli di dettaglio territoriale, per un totale di 143 vincoli imposti a livello dell'intero campione e per il solo quarto entrante.

Le informazioni della BDR sono qui utilizzate per introdurre 18 nuovi vincoli, corrispondenti al numero dei dichiaranti di reddito fiscale per le tre principali tipologie di reddito e per i tre gruppi (reddito basso, medio, alto), sia per l'intera popolazione, sia per il quarto entrante. In questo modo ci assicuriamo che il campione riproduca la "struttura" della distribuzione dei dichiaranti della BDR contenuta nella Tavola 12.6.

12.6 L'impatto dei nuovi coefficienti di riporto all'universo sulle stime finali

Se il nuovo sistema di ponderazione inclusivo delle variabili provenienti da BDR assicura la conformità del campione EU-SILC alla struttura dei dichiaranti di reddito della BDR, qual è l'effetto sulle stime campionarie relative agli aggregati di reddito fiscale della BDR e alle principali variabili di interesse dell'indagine?

La tavola 12.9 mostra le stime campionarie dei principati aggregati della BDR, e va dunque confrontata con la precedente tavola 12.3 per valutare la capacità dei nuovi pesi di rappresentare il reddito fiscale. La differenza tra BDR e EU-SILC relativa al reddito fiscale complessivo è in media dimezzata, e la distanza massima tra i due aggregati è inferiore al 2% nel periodo considerato (Figura 12.1). L'impatto dei nuovi pesi sulle principali tipologie reddituali differisce però in modo significativo. La distanza della stima campionaria dal valore BDR aumenta lievemente per il reddito da lavoro dipendente, a causa del marcato au-

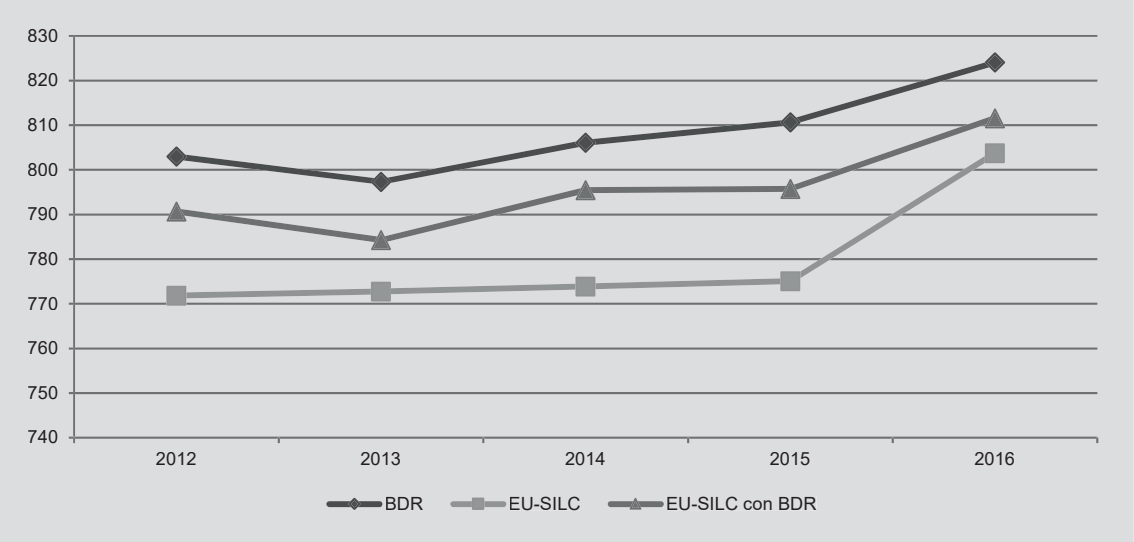
7 Genesees a sua volta implementa la metodologia di calibrazione illustrata in Deville e Särndal (1992).

mento dei dichiaranti a basso reddito: la sottostima dell'aggregato arriva fino a circa 2 punti percentuali. La distanza della stima campionaria dei redditi da lavoro autonomo dal totale BDR è invece drasticamente ridotta, in media di circa i 2/3: la differenze massima tra i due aggregati passa dal 17% al 5% (Figura 12.2). Anche la stima del reddito fiscale da pensione migliora nettamente, e la distanza massima dal totale BDR è inferiore al punto percentuale lungo tutto il periodo considerato.

Tavola 12.9 - Confronto degli aggregati di reddito fiscale per tipologia, da BDR e da stime campionarie EU-SILC dopo la correzione - Totale nazionale (MLN di euro)

		2012	2013	2014	2015	2016
Reddito totale	BDR	802.987	797.320	806.091	810.642	824.046
	EU-SILC	790.680	784.276	795.492	795.753	811.583
	Differenza %	-1,53	-1,64	-1,31	-1,84	-1,51
Reddito da lavoro dipendente	BDR	432.130	429.907	427.193	426.271	433.825
	EU-SILC	424.426	424.519	424.009	418.560	425.741
	Differenza %	-1,78	-1,25	-0,75	-1,81	-1,86
Reddito da lavoro autonomo	BDR	106.142	107.713	104.957	105.198	106.302
	EU-SILC	103.626	102.829	99.752	101.080	102.400
	Differenza %	-2,37	-4,53	-4,96	-3,91	-3,67
Reddito da pensione	BDR	233.864	238.810	243.617	247.201	249.218
	EU-SILC	233.673	238.755	244.651	247.436	250.608
	Differenza %	-0,08	-0,02	0,42	0,10	0,56

Figura 12.1 - Reddito complessivo fiscale - Aggregato BDR e stima campionaria EU-SILC con e senza i nuovi coefficienti di riporto all'universo (miliardi di euro)



La tavola 12.10 mostra invece le stime dei redditi familiari netti dell'indagine EU-SILC ottenute con e senza l'inclusione delle variabili BDR nel calcolo dei coefficienti di riporto all'universo, e dunque permette di valutare l'impatto del nuovo sistema di ponderazione sul reddito delle famiglie. Il reddito familiare cresce in modo significativo in tutto il periodo considerato, con incrementi che vanno dall'1% fino al 2,5%: l'uso delle variabili BDR ha come effetto generale un incremento del peso dei dichiaranti di reddito rispetto ai non dichiaranti che, data la correlazione tra redditi fiscali e redditi di indagine, si traduce in un

incremento del peso dei percettori di reddito. L'impatto complessivo però è la risultate di effetti eterogenei tra le varie tipologie reddituali (Figura 12.3). Il reddito medio familiare da lavoro dipendente diminuisce leggermente, probabilmente a causa dell'incremento del peso dei percettori di redditi fiscali bassi. Al contrario i redditi familiari da lavoro autonomo crescono in modo notevole lungo tutto il periodo, con percentuali che vanno dal 3,5% al 9%: in questo caso all'incremento del peso dei dichiaranti in generale si aggiunge la correzione della sottostima dei percettori di redditi autonomi. Infine, anche i redditi familiari da pensione crescono, con percentuali che variano dal 2,5% al 4,5%.

Figura 12.2 - Reddito da lavoro autonomo fiscale - Aggregato BDR e stima campionaria EU-SILC con e senza i nuovi coefficienti di riporto all'universo (miliardi di euro)

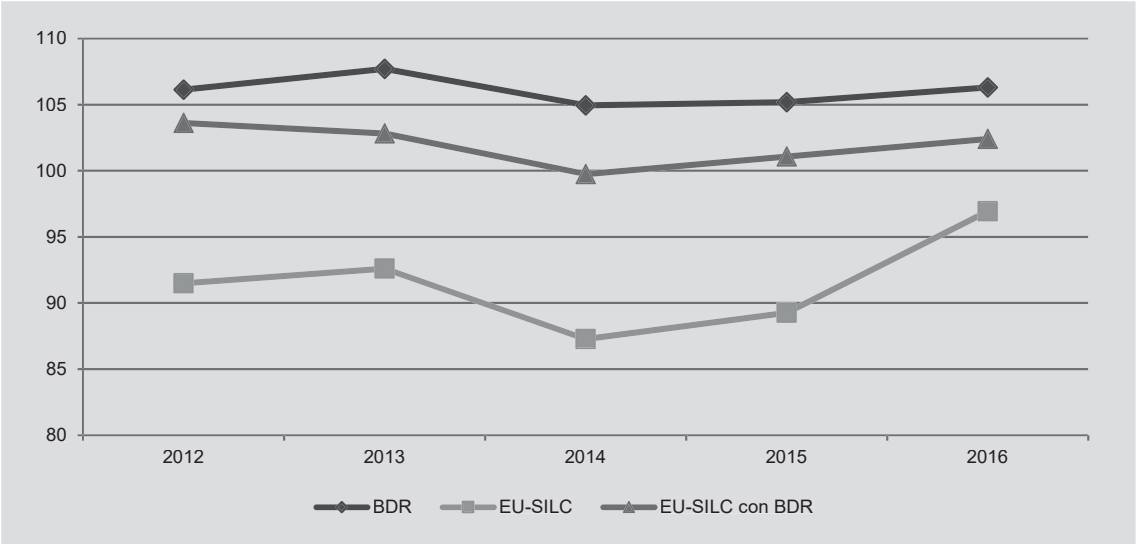


Tavola 12.10 - Confronto dei redditi familiari netti per tipologia. Stime campionarie EU-SILC con e senza i nuovi coefficienti di riporto all'universo. Anni 2012-2016 (valori medi in euro e variazioni percentuali)

		2012	2013	2014	2015	2016
Reddito totale	BDR	30.236	29.579	29.473	29.472	29.987
	EU-SILC	30.808	30.019	30.215	30.056	30.284
	Differenza %	1,89	1,49	2,52	1,98	0,99
Reddito da lavoro dipendente	BDR	13.598	13.479	13.420	13.706	13.831
	EU-SILC	13.523	13.277	13.318	13.532	13.551
	Differenza %	-0,55	-1,50	-0,76	-1,27	-2,02
Reddito da lavoro autonomo	BDR	5.365	5.106	4.871	4.643	4.860
	EU-SILC	5.785	5.506	5.340	5.007	5.033
	Differenza %	7,83	7,83	9,63	7,84	3,56
Reddito da pensione	BDR	8.764	8.769	8.733	8.724	8.778
	EU-SILC	8.988	8.991	9.037	9.119	9.173
	Differenza %	2,56	2,53	3,48	4,53	4,50

Figura 12.3 - Redditi familiari netti per tipologia. Stime campionarie EU-SILC con e senza i nuovi coefficienti di riporto all'universo. Anni 2012-2016 (valori medi in euro)

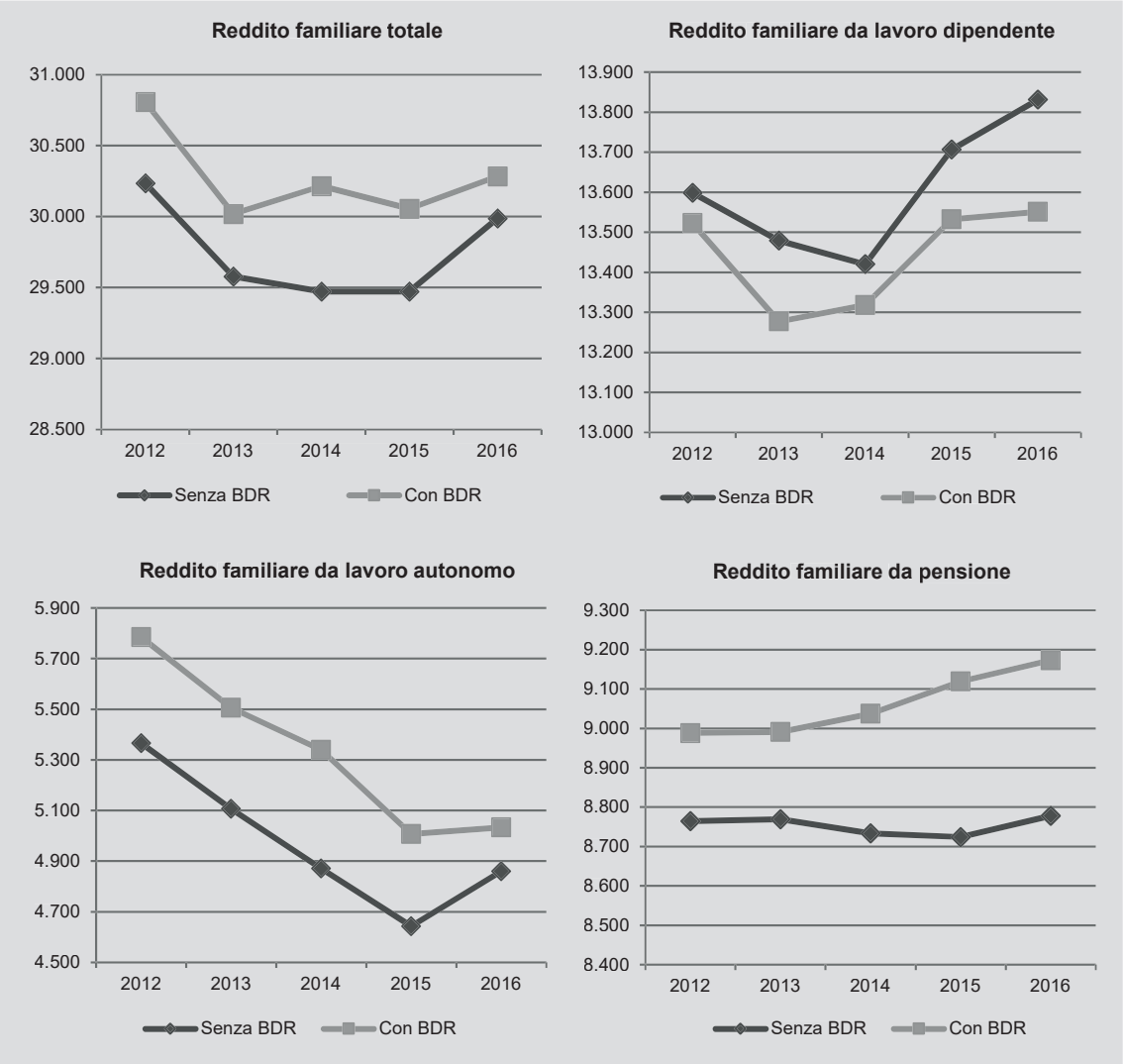


Tavola 12.11 - Principali indicatori di povertà, esclusione sociale e disuguaglianza dell'indagine EU-SILC con e senza i nuovi coefficienti di riporto all'universo. Anni 2012-2016 (valori percentuali e differenze)

		2012	2013	2014	2015	2016
Rischio di povertà o esclusione sociale	Pesi senza BDR	19,32	19,26	19,66	19,98	20,72
	Pesi con BDR	19,50	19,29	19,45	19,93	20,63
	Differenza	-0,18	-0,03	0,21	0,05	0,09
Grave Deprivazione materiale	Pesi senza BDR	14,19	12,18	11,36	11,14	11,94
	Pesi con BDR	14,45	12,34	11,60	11,47	12,08
	Differenza	-0,26	-0,16	-0,24	-0,33	-0,14
Indice di Gini	Pesi senza BDR	32,44	32,83	32,40	32,40	33,15
	Pesi con BDR	32,59	32,91	33,02	32,58	32,99
	Differenza	-0,14	-0,07	-0,62	-0,18	0,16

La tavola 12.11 infine, mostra l'impatto della ripesatura sui principali indicatori dell'indagine EU-SILC: le variazioni del rischio di povertà o esclusione sociale, dell'indicatore di grave deprivazione materiale e dell'indice di Gini sono contenute e mai statisticamente significative. L'impatto distributivo della procedura di ponderazione inclusiva delle variabili BDR sembra dunque piuttosto limitato.

12.7 Conclusioni

L'inclusione delle informazioni provenienti dalla Banca Dati Reddituali nel campione di EU-SILC permette di valutare l'ampiezza della distorsione campionaria relativa alle variabili fiscali: appare evidente una generale sottostima tanto del numero dei dichiaranti che dei principali aggregati di reddito fiscale. Tale distorsione si presenta inoltre come eterogenea, sia rispetto alle diverse tipologie di reddito fiscale considerate sia relativamente al livello del reddito dei contribuenti. La sottostima dei redditi fiscali da lavoro autonomo appare più marcata di quella dei redditi da lavoro dipendente o da pensione: per i primi la distanza dal corrispondente aggregato BDR arriva fino al 17% circa nel periodo considerato, per i secondi varia tra l'1,5% e il 4%. Per i redditi fiscali da lavoro autonomo, inoltre, sono i dichiaranti appartenenti all'ultimo quartile di reddito fiscali a risultare stimati con minore precisione, mentre l'opposto accade per i redditi dipendenti o pensionistici, per i quali la distorsione maggiore riguarda i dichiaranti a basso reddito. La strategia di contenimento della distorsione è dunque relativamente complessa e utilizza le informazioni della BDR relative tanto alle diverse tipologie reddituali che alla posizione dei dichiaranti nella distribuzione del reddito fiscale. La procedura di costruzione dei coefficienti di riporto all'universo di EU-SILC è stata quindi modificata attraverso l'uso delle variabili BDR tanto nella correzione del peso diretto per la probabilità di mancata risposta che per la costruzione di 18 nuovi totali noti relativi alla struttura della "popolazione fiscale" cui calibrare il campione. L'impatto della nuova procedura di costruzione dei pesi sulle stime campionarie dei principali aggregati fiscali è notevole: la distanza tra stime e totali BDR si riduce notevolmente, dimezzandosi per il reddito fiscale complessivo e riducendosi circa dei 2/3 per i redditi da lavoro autonomo. I nuovi pesi hanno inoltre un impatto importante sui redditi familiari netti, la principale variabile dell'indagine EU-SILC: il reddito familiare totale cresce in modo significativo lungo tutto il periodo (tra l'1% e il 2,5%), e in particolare i redditi familiari da lavoro autonomo risultano molto più elevati (incrementi che vanno dal 3,5% al 9%). L'effetto sui principali indicatori di povertà, esclusione sociale e disuguaglianza appare invece contenuto: sembra dunque che l'impatto distributivo della procedura di ponderazione inclusiva delle variabili BDR sia piuttosto limitato.

PARTE QUARTA

13. IL TRATTAMENTO E L'UTILIZZO DEI DATI LONGITUDINALI[1]

13.1 Introduzione

L'indagine EU-SILC fornisce un ricco contributo informativo, che presenta il notevole vantaggio di essere rilevato di anno in anno sulle medesime unità di rilevazione, per un periodo di osservazione che può durare fino a quattro anni consecutivi. Ciò consente di osservare in particolare i cambiamenti vissuti dagli individui e dalle loro famiglie e di analizzare come i diversi percorsi biografici (lavorativi, familiari, procreativi, etc.) si influenzino a vicenda e contribuiscano a determinare specifiche condizioni economiche. Ad esempio, oltre a descrivere le caratteristiche degli individui a rischio di povertà o esclusione sociale, l'indagine consente di comprendere quali caratteristiche e quali percorsi portino a un maggior rischio di entrare, uscire o permanere in tale condizione. È fondamentale avere la possibilità di integrare la lettura trasversale dei fenomeni con quella longitudinale, per comprendere i meccanismi che contribuiscono a determinare le condizioni di disagio degli individui e delle loro famiglie e potere così identificare le popolazioni più fragili e i comportamenti di rischio più rilevanti, cui indirizzare eventuali politiche sociali.

Per questo motivo, Eurostat richiede che i diversi Paesi adottino delle strategie di rilevazione che consentano di produrre stime sia trasversali sia longitudinali. Tra gli indicatori longitudinali principali vi è la persistenza a rischio di povertà, che misura la percentuale di individui a rischio di povertà in un determinato anno e in almeno due dei tre anni precedenti. Per il calcolo di questo indicatore è quindi necessario avere a disposizione almeno 4 osservazioni consecutive sulle stesse unità statistiche. Di conseguenza, Eurostat raccomanda l'uso di un disegno campionario basato su un panel ruotato composto da quattro gruppi rotazionali. In linea di principio, sarebbe possibile usare disegni campionari indipendenti per le stime trasversali e quelle longitudinali oppure usare un panel integrato composto da più di quattro gruppi rotazionali. Tuttavia, l'Istat, come la maggior parte dei paesi che partecipano ad EU-SILC, ha adottato il disegno raccomandato da Eurostat, in quanto rappresenta un giusto compromesso tra le esigenze di produrre stime longitudinali, relative alle dinamiche osservabili tra 2, 3 e 4 anni consecutivi, e stime trasversali basate su campioni che siano il più possibile rappresentativi della popolazione dell'anno di rilevazione (un quarto del campione teorico è formato infatti dal campione estratto dalla popolazione dell'anno di rilevazione, ma ben maggiore è il suo peso relativo nel campione effettivo, per via delle mancate risposte osservate nei panel estratti negli anni precedenti).

1 Il capitolo è stato redatto da Lucia Coppola.

13.2 Unità di rilevazione longitudinale e regole di inseguimento

Per specificare i campioni longitudinali è necessario definire la popolazione oggetto di studio, le unità che la rappresentano e le strategie di inseguimento e rilevazione durante il periodo di osservazione del panel. Tali definizioni sono state formalizzate nel regolamento Europeo e sono pertanto vincolanti per la rilevazione a livello nazionale (European Commission, 2003).

La popolazione oggetto di studio, come detto in precedenza, è rappresentata dagli individui residenti in famiglia sul territorio nazionale. Quindi se un individuo emigra al di fuori del territorio nazionale o si trasferisce in istituzione non fa più parte della popolazione oggetto di indagine e non è più necessario che sia inseguito e intervistato. Invece i trasferimenti dell'intera famiglia o dei componenti familiari definiti "campione" nell'ambito del territorio nazionale danno luogo all'inseguimento e all'intervista presso il nuovo indirizzo di residenza.

Poiché la composizione delle famiglie è soggetta a cambiamenti nel tempo, in quanto tra un anno e l'altro alcuni componenti possono uscire dalla famiglia e altri entrare a farne parte, è difficile definire la famiglia come unità di rilevazione longitudinale. Per semplicità si fa quindi riferimento agli individui e si richiede di rilevare i cosiddetti individui campione e le loro famiglie. Questa strategia, oltre ad essere di più semplice attuazione, consente di rappresentare le dinamiche di formazione e scioglimento delle famiglie e come queste si relazionino alle condizioni di vita. Si rende però necessario definire esattamente quali individui debbano essere inseguiti e rilevati nel tempo.

In linea di principio bisognerebbe reintervistare tutti i componenti familiari rilevati nella prima occasione di indagine. Tuttavia, al fine di contenere il carico sia sui rispondenti sia sugli istituti nazionali di statistica, Eurostat definisce individui campione tutti i componenti familiari di almeno 14 anni, rilevati nella prima occasione di indagine[2]. Sono questi gli individui che devono essere reintervistati nei 3 anni successivi, anche se cambiano abitazione e lasciano la famiglia di origine. In questo caso daranno vita ad una nuova famiglia campione (denominata famiglia split), in cui tutti i componenti familiari eleggibili devono essere intervistati. Per quanto riguarda i componenti familiari che alla prima intervista hanno meno di 14 anni, questi vengono definiti coresidenti e non devono essere inseguiti qualora dovessero cambiare abitazione nell'ambito del territorio nazionale. Nella maggior parte dei casi, i trasferimenti di questi individui avvengono assieme ad altri individui campione, per cui di fatto vengono inseguiti e rilevati come componenti delle famiglie degli individui campione. Nel caso in cui si trasferiscano senza altri individui campione, invece, escono dalla rilevazione. Analogamente, gli individui che entrano a far parte delle famiglie di un individuo campione dalla seconda intervista in poi vengono definiti coresidenti e non devono essere più inseguiti e rilevati se si trasferiscono di nuovo presso famiglie che non contengono almeno un individuo campione.

Una volta definiti gli individui campione in occasione della prima rilevazione, si devono definire le regole secondo cui questi debbano essere inseguiti nell'ambito del territorio nazionale, a seconda dell'esito dei tentativi di contatto nelle rilevazioni successive. È possibile, infatti, che durante il periodo di osservazione il contatto con le famiglie degli individui campione non abbia successo o che non si possa procedere con l'intervista. A seconda delle

2 Il regolamento definisce l'età minima al di sopra della quale un componente familiare debba essere definito "campione". Gli istituti nazionali di statistica possono quindi scegliere di utilizzare una soglia inferiore ma non superiore. L'adozione di una soglia inferiore è raccomandata in caso di panel di durata superiore ai 4 anni.

motivazioni della mancata intervista e della loro frequenza durante i quattro anni teorici di osservazione, si rende necessario definire delle strategie di inseguimento ragionevoli, in modo tale da contenere il carico sia sui rispondenti sia sugli istituti nazionali di statistica. Tali regole, definite sempre tramite regolamento europeo, stabiliscono che se una famiglia rifiuta di collaborare, o se l'indirizzo di residenza risulta non esistente o non raggiungibile, non si debba procedere con un ulteriore tentativo di contatto nel successivo anno di indagine. Se invece la mancata intervista è dovuta a motivi temporanei, quali l'assenza o l'impossibilità di partecipare all'intervista (ad esempio per motivi di salute) durante il periodo di rilevazione, si deve tentare di contattare nuovamente la famiglia nella successiva occasione di indagine. Tuttavia, se questa tipologia di mancata risposta si dovesse verificare alla prima intervista o per due anni consecutivi, non è necessario procedere con un ulteriore tentativo di contatto. Queste regole implicano che durante il periodo di osservazione del panel si presentino casistiche in cui l'informazione sulle famiglie degli individui campione siano talvolta mancanti per uno o più anni, in modo anche discontinuo.

13.3 Principali criticità della rilevazione longitudinale

La rilevazione longitudinale fornisce indubbiamente un contenuto informativo prezioso. Tuttavia, la rilevazione di tale contenuto presenta delle criticità specifiche poiché è necessario predisporre delle procedure, sia in fase di rilevazione sia in fase di trattamento dati, atte a garantire la coerenza nel tempo delle informazioni rilevate.

In primo luogo, data la complessa articolazione delle regole di inseguimento illustrate nel paragrafo precedente, bisogna garantire la corretta identificazione degli individui campione e degli altri componenti familiari durante tutta la durata del panel. A questo fine, non basta assegnare una chiave univoca a ciascuno di loro in fase di rilevazione, ma si rende necessario predisporre delle procedure di controllo e correzione che garantiscano che allo stesso individuo venga correttamente assegnata la medesima chiave identificativa durante tutta la durata del panel e che vengano generate opportunamente le nuove chiavi identificative per i componenti familiari che entrano nella rilevazione successivamente alla prima intervista. Data la dinamica che caratterizza la composizione delle famiglie, per cui alcuni componenti possono uscire e rientrare, formare nuove famiglie e tornare nella famiglia di origine, la soluzione a questo tipo di problema può essere complessa e onerosa. Per questo motivo è stata predisposta una procedura generalizzata che viene illustrata e discussa nel capitolo 14 "La trasformazione dei file provenienti dalla rilevazione alla base dei processi di trattamento dei dati dell'indagine".

Una volta identificate e rilevate correttamente le unità di rilevazione longitudinali, bisogna considerare che le informazioni rilevate tramite interviste svolte in anni successivi possono entrare in contraddizione, per via delle diverse tipologie di errori non campionari che si possono verificare nei vari anni: si pensi ad esempio ad una errata codifica da parte del rilevatore, ad errori di memoria o di interpretazione del quesito da parte del rispondente, e ad altri errori dovuti alla somministrazione del questionario tramite intervista proxy. Al fine di ridurre le incoerenze derivanti dalla fase di rilevazione, si sfruttano le potenzialità del questionario elettronico, proponendo all'intervistato delle domande a conferma su alcune informazioni rilevate nelle interviste precedenti. Questa strategia consente non solo di rendere più agevole la somministrazione del questionario, ma anche di correggere già in fase di rilevazione l'informazione rilevata nelle interviste precedenti. L'articolazione delle

strategie di domande a conferma e gli effetti che l'introduzione di questi quesiti ha avuto sui dati rilevati nella seconda o successiva occasione di indagine vengono discussi nel capitolo 15 "L'impatto della rilevazione proattiva delle informazioni sulla componente longitudinale".

Nonostante l'utilizzo già in fase di rilevazione delle informazioni rilevate nelle occasioni di indagine precedenti, l'analisi longitudinale dei microdati rivela la permanenza di incoerenze tra le informazioni rilevate nelle varie occasioni di indagine, che si ritiene opportuno trattare e correggere prima del rilascio. Le principali strategie di controllo e correzione vengono presentate e discusse nel capitolo 16 "Il trattamento dei dati longitudinali".

Infine, bisogna tener presente che il contributo informativo prodotto dalla rilevazione longitudinale si riferisce alle caratteristiche delle diverse popolazioni longitudinali che possono essere rappresentate sia dai singoli panel sia da possibili combinazioni dei panel. Per questo motivo viene predisposto un sistema di pesi che, a partire dalle unità di rilevazione effettivamente intervistate di anno in anno (al netto quindi delle mancate risposte che possono introdurre problemi di selezione sul campione), consentano di sfruttare pienamente le informazioni longitudinali in modo flessibile a seconda degli obbiettivi conoscitivi. Il capitolo 17 "L'utilizzo dei dati in ottica longitudinale: potenzialità del disegno campionario a gruppi rotazionali" illustra le caratteristiche dei panel, i percorsi di osservazione delle unità di rilevazione e le diverse strategie di analisi che possono essere adottate tramite l'utilizzo del sistema di pesi prodotto.

14. LA TRASFORMAZIONE DEI FILE PROVENIENTI DALLA RILEVAZIONE ALLA BASE DEI PROCESSI DI TRATTAMENTO DEI DATI DELL'INDAGINE[1]

14.1 Introduzione

Il processo di produzione dell'informazione statistica derivante dai dati dell'indagine europea sul reddito e le condizioni di vita delle famiglie italiane (EU-SILC) è articolato in diverse fasi di lavorazione. Il presente capitolo descrive la fase iniziale del processo annuale di produzione dei dati, che già in fase di acquisizione dei dati di rilevazione implementa le procedure per il controllo delle chiavi individuali e familiari, la correzione delle errate assegnazioni e la generazione delle nuove chiavi, tenendo conto della natura longitudinale della rilevazione.

Essenzialmente, il requisito software da soddisfare è quello di rendere efficiente la gestione degli eventuali cambiamenti apportati alla rilevazione, in termini di contributo informativo rilevato, quesiti somministrati e quindi variabili da trattare. La manutenzione deve richiedere il minor sforzo possibile da parte del personale coinvolto in tale fase. Il software realizzato è stato scritto in SAS System (installazione base), che rappresenta il pacchetto statistico utilizzato per il trattamento dei dati EU-SILC. Poiché l'approccio adottato si può considerare "generalizzato", necessita di poca manutenzione e richiede un impegno contenuto da parte di un solo incaricato.

La rilevazione sul campo e l'acquisizione dei dati dell'indagine EU-SILC è gestita da una ditta esterna che rilascia all'Istat dei file in formato ASCII, contenenti i dati raccolti tramite la somministrazione del questionario elettronico. La procedura in esame si occupa, in primo luogo, del caricamento dei file suddetti e della loro trasformazione in un formato idoneo alle procedure di controllo e correzione dei dati. Si fa uso, in modo soddisfacente, dei metadati derivanti dai modelli di rilevazione, indipendentemente dalla tecnica di rilevazione usata. Il maggior numero delle variabili contenute nei vari data set sono create in questa fase. Il software realizzato controlla che queste variabili siano conformi alle specifiche dettate alla ditta esterna responsabile del software di acquisizione dei dati. Inoltre, la procedura controlla l'integrità referenziale dei dati, cioè ogni data set prodotto deve essere coerente con l'unità di rilevazione al quale si riferisce.

La natura longitudinale dell'indagine EU-SILC determina la necessità di controlli rigorosi per l'attribuzione degli identificativi delle osservazioni dei vari data set. A tale scopo, è stato realizzato un algoritmo per la valorizzazione delle chiavi da attribuire a tutti i record, con particolare attenzione agli individui che entrano a far parte della famiglia a partire dalla seconda intervista. Inoltre, le famiglie derivanti da trasferimenti di individui campione nell'ambito del territorio nazionale (famiglie *split*) e quelle nate dalla fusione di due o più famiglie già rilevate in precedenza (famiglie *fusion*), determinano un aumento della complessità della fase di attribuzione delle chiavi identificative. Infine, è necessario identificare correttamente gli individui rientrati nel campione dopo un periodo di assenza temporanea,

1 I paragrafi 14.1, 14.2 e 14.6 sono stati redatti da Pierpaolo Massoli; i paragrafi 14.3, 14.4 e 14.5 sono stati redatti da Giovanni Battista Arcieri.

distinguendoli da coloro che entrano nel campione per la prima volta, in modo tale da garantire che mantengano la chiave identificativa attribuita loro in occasione della prima intervista.

Il capitolo è articolato come segue: il Paragrafo 14.2 descrive il modello relazionale dei dati dell'indagine; il Paragrafo 14.3 descrive brevemente il software realizzato per la trasformazione dei dati ASCII in data set SAS ed il sistema di metadati utilizzato per guidare il processo di controllo e correzione dei dati dell'indagine; il Paragrafo 14.4 pone l'attenzione sull'algoritmo di creazione degli identificativi delle osservazioni nei data set; il Paragrafo 14.5 riporta la soluzione adottata in EU-SILC per risolvere il problema della corretta attribuzione degli identificativi agli individui che escono temporaneamente dall'indagine per poi rientrarvi; alcune considerazioni critiche sono conclusivamente riportate nel Paragrafo 14.6.

14.2 L'organizzazione dei dati

Nell'indagine EU-SILC le informazioni provenienti dalle interviste sono organizzate in quattro distinti data set, secondo l'unità di rilevazione di riferimento: (1) la scheda contatti, (2) la scheda famigliare, (3) il questionario famigliare e (4) il questionario individuale.

Il primo data set contiene tutte le informazioni necessarie per contattare la famiglia, come i nominativi dei componenti della famiglia anagrafica, l'indirizzo e i recapiti telefonici della famiglia, i tentativi e gli esiti di contatto del rilevatore con la famiglia. Il data set è quindi composto da un record per ogni famiglia del campione teorico. Il secondo data set riguarda le informazioni rilevate per tutti i componenti famigliari, inclusi quelli non eleggibili per l'intervista individuale, come ad esempio le caratteristiche anagrafiche, i motivi dell'assenza di un componente, ecc. In questo data set sono considerati anche gli individui di fatto appartenenti alla famiglia cioè coloro che pur non facendo parte della famiglia anagrafica, sono considerati componenti familiari secondo la definizione adottata[2]. Il data set, quindi, è composto da un record per ogni componente famigliare, che sia presente o meno nella famiglia al momento dell'intervista. Inoltre, nel data set è presente anche un record per ogni famiglia che, pur appartenendo al campione teorico, non è stata intervista con successo, con le informazioni relative al motivo per cui il contatto con la famiglia o l'intervista non hanno avuto luogo. Il data set famigliare è costituito da un record per ogni famiglia effettivamente intervistata, e contiene tutte le informazioni rilevate a livello familiare, come ad esempio le caratteristiche e il titolo di godimento dell'abitazione, le condizioni economiche della famiglia ecc. Infine, il data set individuale è costituito da un record per ogni componente familiare eleggibile per l'intervista individuale (cioè che abbia compiuto almeno 16 anni) e contiene le informazioni rilevate a livello individuale, come le condizioni di salute, l'attività lavorativa svolta, le componenti di reddito percepite, ecc.

Ogni record nei vari data set è individuato da una chiave identificativa composta da uno o più attributi. La chiave identificativa famigliare (*DB030*) è un numero intero progressivo e quella individuale è un numero composto dall'unione della chiave famigliare e un numero intero progressivo (denominata *RB030* nella scheda famigliare e *PB030* nel questionario individuale). Ogni famiglia che entra a far parte dell'indagine viene identificata con una chiave famigliare che mantiene per quattro anni, a meno che non esca dall'indagine secondo le

2 La definizione di famiglia adottata dall'Istat per la rilevazione EU-SILC è quella della famiglia di fatto, ovvero l'insieme delle persone che vivono abitualmente - attualmente presenti o temporaneamente assenti - nella stessa abitazione, legate da vincoli di parentela, affinità, adozione, tutela, affetto o amicizia.

regole di inseguimento definite da regolamento (European Commission, 2003). Analogamente, ogni individuo viene identificato con una chiave che mantiene anche qualora uscisse dalla famiglia o ne formasse una nuova (la sua chiave non può essere attribuita ad un altro componente della famiglia).

Data la dinamica longitudinale di EU-SILC, uno o più individui possono uscire dalla famiglia d'origine e formare una nuova famiglia (famiglia *split*). D'altra parte, i componenti di due o più famiglie intervistate si possono unire formando una singola famiglia (famiglia *fusion*). Queste situazioni richiedono estrema cura nell'attribuzione dei codici identificativi sia famigliari che individuali.

I record della scheda famigliare sono identificati dalla coppia (*DB030*, *RB030*) e rappresentano ciascun individuo (*RB030*) facente parte della famiglia (*DB030*). I record del questionario famigliare sono identificati dalla sola chiave *DB030*, mentre quelli del questionario individuale sono identificati dalla chiave *PB030*. Questi identificativi sono vincolati tra loro, in una sorta di integrità referenziale, cioè non possono esistere dei questionari individuali se non esiste un questionario famigliare di riferimento e gli individui non sono registrati nella scheda famigliare. Tale aspetto è già rilevante in una indagine di natura trasversale, e lo diventa ancor di più in una di natura longitudinale, in cui si rende necessario preservare l'integrità referenziale tra le unità di rilevazione nel tempo, tenendo conto delle dinamiche familiari che possono aver luogo secondo le regole di inseguimento implementate.

14.3 Il software di trasformazione dei dati

Nell'ambito dell'indagine EU-SILC, ogni anno il processo di controllo e correzione dei dati si articola essenzialmente in dieci fasi di lavorazione sequenziali, per cui i dati di output di ogni fase rappresentano l'input per la fase successiva. È possibile che una medesima fase debba essere eseguita più volte. Per tenerne traccia, in ogni data set è presente una variabile *VERSFASE*, che può assumere i valori 0, -1, -2, ... e viene aggiornata ad ogni esecuzione, in modo tale che la modalità 0 della variabile identifichi sempre la versione più aggiornata dei dati (ovvero l'ultima esecuzione).

La ditta che si occupa della rilevazione sul campo garantisce la fornitura dei dati rilevati tramite un unico file di testo in formato ASCII, secondo le specifiche concordate con gli esperti dell'indagine. Il processo di controllo e correzione richiede invece che i dati siano in formato SAS e quindi la fase di trasformazione dei dati assume un ruolo tutt'altro che secondario. Il software realizzato a tal fine è costituito da una serie di programmi scritti in linguaggio SAS Macro che è compreso nel pacchetto base del SAS System. La scelta del linguaggio macro ha permesso di scrivere dei programmi sufficientemente generalizzati da richiedere il minor impegno possibile per la loro manutenzione. Per raggiungere questo scopo vengono sfruttati i metadati memorizzati in appositi file esterni al software, ovvero dei fogli di lavoro di MS Excel, nei quali sono riportati: i nomi delle variabili del file unico fornito dalla ditta esterna e i nomi delle variabili adottati nei data set SAS, il tipo della variabile (testo o numerica), il campo di validità, a quale data set la variabile appartiene e se la variabile deve essere tenuta nel data set o meno. Il software è scritto in modo da poter gestire tutti i cambiamenti apportati ai modelli di rilevazione, secondo quanto specificato nei file di metadati, senza dover essere modificato a sua volta. Il file unico è quindi importato dal server della ditta esterna e trasformato nei quattro data set SAS standard del processo di controllo e correzione.

A seguito dell'importazione dei dati di rilevazione, viene verificata la correttezza degli identificativi famigliari (*DB030*) ed individuali (*RB030*), accertando in primo luogo che non vi siano valori duplicati delle chiavi. I componenti aggiuntivi di fatto della famiglia, che non siano quindi presenti nella scheda anagrafica prima dell'intervista e che non siano già stati rilevati negli anni precedenti, sono importati nei data set con una chiave individuale costruita sequenzialmente a partire dall'ultimo identificativo individuale valorizzato per la famiglia di appartenenza, in base ad un algoritmo che viene descritto dettagliatamente nel Paragrafo 14.4. Tale chiave rimane comunque provvisoria fintanto che non viene eseguito anche l'algoritmo descritto nel Paragrafo 14.5, atto a distinguere i nuovi componenti familiari da quelli "rientrati", ovvero quei componenti familiari già rilevati in precedenza, che risultano usciti dalla famiglia in un'occasione di indagine e rientrano a farvi parte ad una intervista successiva. Ai nuovi componenti, infatti, è necessario attribuire una nuova chiave identificativa, mentre ai "rientrati" è necessario riassegnare correttamente la loro chiave identificativa originale.

Dopo l'importazione del file ASCII in data set SAS, il programma verifica la coerenza della scheda famigliare e del questionario famigliare, cioè verifica l'esistenza di un questionario famigliare a seconda dell'esito dell'intervista riportato sulla scheda famigliare. Analogamente, si procede alla verifica della coerenza del questionario famigliare e dei questionari individuali. Vengono aggiunti al data set della scheda famigliare i record delle famiglie non rispondenti (che si identificano con la *DB030* della famiglia cui viene associato una chiave individuale fittizia *RB030* = 99). Inoltre, si aggiunge un record per ciascuna famiglia che alla fine del periodo di rilevazione risulta non essere stata contattata (registri d'ufficio). Ciò consente di avere tutte le informazioni necessarie per l'analisi delle cadute totali delle interviste e valutare la performance della tecnica di rilevazione adottata, l'efficienza della rete di rilevazione, valutazioni dell'*attrition* per la componente longitudinale, ecc. Infine, si creano diverse variabili che non sono direttamente rilevate tramite questionario ma che sono utili per il controllo e la correzione delle fasi successive di lavorazione. Sono essenzialmente variabili di natura geografica e di disegno campionario.

14.4 L'algoritmo di creazione degli identificativi individuali

A meno di uscite dall'indagine per vari motivi, una famiglia viene coinvolta per un periodo di quattro anni (panel costituito da quattro *wave*). Il campione dell'indagine EU-SILC si basa su un campione a quattro gruppi rotazionali, per cui ogni anno viene estratto un nuovo campione che entra in rilevazione in sostituzione di quello già intervistato per quattro volte. Lo schema che riporta tale dinamica, ad esempio dal 2004 al 2010, è mostrato in Tavola 14.1.

Tavola 14.1 - Schema di campionamento di EU-SILC per gli anni dal 2004 al 2010

CAMPIONE	Anno						
	2004	2005	2006	2007	2008	2009	2010
C1	w4						
C2	w3	w4					
C3	w2	w3	w4				
C4	w1	w2	w3	w4			
C5		w1	w2	w3	w4		
C6			w1	w2	w3	w4	
C7				w1	w2	w3	w4

Nel 2004 (anno di inizio dell'indagine), tutti e quattro i campioni longitudinali (contrassegnati nello schema con C1, C2, C3 e C4) hanno partecipato all'indagine per la prima volta. Per iniziare la rotazione, si è ipotizzato che il campione C1 fosse alla quarta *wave* (w_4 nello schema), il campione C2 fosse alla terza *wave* (w_3), il campione C3 alla seconda (w_2). Quello indicato con C4 è il primo campione longitudinale che, iniziato nel 2004, proseguirà per 4 *wave*, come da disegno, e consentirà la realizzazione del primo campione longitudinale completo (composto da w_1, w_2, w_3, w_4). Nel 2005, C5 è il nuovo campione longitudinale che entra nell'indagine e ha preso il posto di C1 terminato nel 2004. In genere, un nuovo campione longitudinale è composto dalle stesse unità di primo stadio (i comuni) e da nuove unità di secondo stadio (le famiglie). Il campione trasversale di uno specifico anno di indagine è composto dall'unione dei quattro campioni longitudinali, ognuno ad una *wave* diversa di rilevazione: in tal modo, in assenza di *attrition*, ogni campione trasversale sarebbe composto da un quarto di famiglie che hanno effettuato la prima intervista, un quarto di famiglie che hanno effettuato la seconda intervista, un quarto di famiglie la terza e infine un quarto di famiglie hanno effettuato la quarta intervista.

La chiave famigliare è composta da una componente fissa (*COD_FAM*) e da una componente variabile rappresentata dalle ultime due cifre (*SPLIT*) in modo che la chiave famigliare: $DB030 = COD\_FAM \times 100 + SPLIT$. Agli inizi dell'indagine EU-SILC, le chiavi identificative delle famiglie del campione (*DB030*) sono state univocamente valorizzate con un *COD_FAM* progressivo da 1 a N e un codice $SPLIT = 00$ (nessun evento *split*). Seguendo lo schema sopra riportato, le famiglie del nuovo campione entrante alla *wave* successiva sono numerate progressivamente con un *COD_FAM* valorizzato a partire da $N+1$ in poi. La valorizzazione delle ultime due cifre della chiave familiare dipende dagli eventi *split* e *fusion*.

Le indagini longitudinali presentano il problema della corretta identificazione degli individui campione che vanno inseguiti per tutte le *wave*, sia che restino nella famiglia cui appartenevano alla prima *wave*, sia che si trasferiscano in altre famiglie campione (*fusion*), sia che formino nuove famiglie (*split*). Le chiavi individuali (*RB030* e *PB030*) sono numerate progressivamente all'interno di ciascuna famiglia secondo la seguente regola: $RB030 = DB030 \times 100 + i$ dove per $i = 1, \ldots, M$ si intende il numero d'ordine dell'i-esimo componente della famiglia di M componenti identificata da quel *DB030*. Si deve evitare di attribuire ad un nuovo soggetto l'identificativo di un altro individuo già intervistato in precedenza ma che non fa più parte della famiglia. Per una corretta costruzione delle chiavi individuali occorre in primo luogo generare correttamente la chiave famigliare tenendo presente tutte le situazioni che vanno a modificare la composizione della famiglia originaria. Pertanto, per una corretta assegnazione dell'identificativo famigliare, è necessario gestire la formazione di famiglie *split* e *fusion*.

- *Generazione di una nuova famiglia a seguito di un evento SPLIT*: l'evento si verifica quando uno o più componenti campione lasciano la famiglia di origine (cioè quella in cui sono stati intervistati il primo anno) per trasferirsi presso un altro domicilio sul territorio nazionale e formano una nuova famiglia. Lo *split* costituisce una suddivisione della famiglia originaria in uno o più famiglie da seguire e intervistare. Non ricadono in questa casistica gli eventuali cambiamento di domicilio dell'intera famiglia, poiché in questi casi la famiglia mantiene lo stesso identificativo. Un esempio di evento *split* è il seguente: la famiglia $DB030 = 100$ ($COD\_FAM = 1$ e $SPLIT = 00$) è composta dai seguenti individui: {10001, 10002, 10003, 10004, 10005}. Supponendo che l'individuo (campione) $RB030 = 10002$ cambi domicilio, egli forma una nuova famiglia identificata da una nuova chiave famigliare *DB030* derivata dalla precedente

aumentando il codice *split* da N_s^{new} a dove $N_s^{new} = N_s^{last} + 1$ con N_s^{last} che indica l'ultimo codice *split* assegnato nella storia delle formazioni di nuove famiglie a partire dalla famiglia con COD_FAM = 1. Se, per esempio, risultasse N_s^{last} = 0 allora la nuova famiglia dell'individuo $RB030$ = 10002 sarebbe $DB030$ = 1 x 100 + N_s^{last} = 101. È importante sottolineare che, l'identificativo $RB030$ dell'individuo rimane comunque sempre lo stesso. Supponendo che, nello stesso tempo, gli individui 10003 e 10005 si trasferiscano insieme ad un nuovo indirizzo, diverso tanto da quello della famiglia originaria ($DB030$ = 100) quanto da quello del componente famigliare della prima famiglia *split* ($DB030$ = 101), essi formeranno una nuova famiglia con $DB030$ = 102.

- *Generazione di una nuova famiglia a seguito di un evento FUSION*: l'evento consiste nella costituzione di una nuova famiglia a partire da due o più individui campione provenienti da due o più famiglie facenti parte dell'indagine. Il criterio di assegnazione dell'identificativo $DB030$ della nuova famiglia deve considerare i domicili delle famiglie originarie. Nel caso in cui la nuova famiglia vada ad abitare all'indirizzo di una delle due famiglie originarie, per identificare la nuova famiglia si userà la chiave famigliare $DB030$ abbinata a coloro che già vivevano in quel domicilio. Se l'indirizzo della nuova famiglia è diverso da entrambi gli indirizzi delle famiglie originarie, la chiave della famiglia sarà $DB030^{new} = \min(DB030_1^{old}, DB030_2^{old})$. Ad esempio, si supponga che la famiglia $DB030$ = 200 costituita dagli individui campione {20001, 20002} e si "fonda" con la famiglia $DB030$ = 300 costituita dall'individuo campione 30001. La nuova famiglia avrà un nuovo codice identificativo famigliare $DB030^{new}$ = 200 se i componenti delle famiglie originarie vivono presso il domicilio della famiglia con codice 200; se invece gli individui andassero ad abitare al domicilio della famiglia con $DB030$ = 300, la famiglia *fusion* avrebbe il codice $DB030^{new}$ = 300. Se l'indirizzo della nuova famiglia è diverso da entrambi gli indirizzi delle due famiglie originarie, il codice identificativo della nuova famiglia sarà $DB030^{new}$ = 200, perché il più basso tra i due codici identificativi di partenza. La nuova famiglia sarà costituita dagli individui 20001, 20002 e 30001, che quindi mantengono il codice identificativo originale. Come secondo esempio, si supponga ora che la famiglia $DB030$ = 200 sia costituita dagli individui {20001, 20002} e che invece la famiglia $DB030$ = 300 sia costituita dagli individui {30001, 30002, 30003} Se solo gli individui {20002, 30003} andassero a costituire una nuova famiglia al di fuori delle rispettive famiglie di origine, allora tale situazione si può descrivere con due eventi *split* seguiti da un evento di *fusion*. Verranno quindi inizialmente assegnate due nuove chiavi famigliari *split* derivate dalle precedenti e solo successivamente verrà assegnata un'unica chiave famigliare a seguito dell'evento *fusion*.

Convalidata la correttezza degli identificativi famigliari, si procede alla verifica e generazione della chiave individuale in caso di nuovi componenti presenti in famiglia al momento dell'intervista. Gli identificativi individuali già utilizzati sono memorizzati in un apposito data set chiamato STORICO_CHIAVI, che quindi contiene per ciascuna famiglia $DB030$, tutti i codici $RB030$ dei componenti della famiglia, compresi quelli che non ne fanno più parte nell'anno di rilevazione in corso. Ad ogni individuo che entra per la prima volta a far parte di una famiglia campione è attribuito un codice individuale provvisorio composto dalla chiave familiare cui si aggiungono gli ultimi due *digit* sempre pari a 99. Questo codice provvisorio viene poi sostituito con un codice $RB030$ definitivo generato in base ai seguenti passi dell'algoritmo proposto:

1. verifica nello storico degli identificativi individuali in ciascuna famiglia, anno per anno a partire dalla prima *wave*, in modo da non attribuire un codice precedentemente utilizzato;
2. calcolo, per ogni nuovo componente trovato, del massimo valore del numero d'ordine i_{max} degli individui all'interno della famiglia. Ad esempio, se la famiglia $DB030$ = 201 è composta dagli individui {20101, 20102, 20003}, il massimo numero d'ordine è i_{max} = 2;
3. per ogni famiglia $DB030$, ad un nuovo componente della famiglia viene assegnata una chiave $RB030$ = $DB030$ x 100 + (i_{max} + 1). Il massimo del numero d'ordine diventa dunque i_{max} = i_{max} + 1 e viene memorizzato nello storico per consentire eventuali ulteriori assegnazioni per altri nuovi individui entrati nella stessa famiglia. Nell'esempio della famiglia $DB030$ = 201 il nuovo individuo è identificato da $RB030$ = 201 x 100 + (2 + 1) = 20103 e il nuovo numero d'ordine diventa i_{max} = 3 e così via per ogni nuovo componente della famiglia;
4. la chiave $RB030$ viene memorizzata in relazione alla famiglia di appartenenza, aggiornando lo storico delle chiavi individuali.

14.5 L'individuazione degli individui "rientrati" nel campione

Un aspetto critico della procedura descritta nel Paragrafo 14.4 è che non assicura la corretta assegnazione delle chiavi individuali $RB030$ nel caso di individui usciti in una data *wave* e poi rientrati nella famiglia in una *wave* successiva. Questo problema è ancor più evidente nella gestione attuale della rilevazione che richiede ad una ditta esterna di effettuare le interviste sul campo raccogliendo i dati in tanti data set quanti sono i rilevatori impegnati per produrre un unico data set che viene rilasciato all'Istat. A tale ditta viene fornito, prima che inizi la rilevazione sul campo, un elenco degli individui campione da intervistare ottenuto in base alle regole di inseguimento applicate agli individui e alle famiglie delle *wave* precedenti. Ad ogni "nuovo" individuo che il rilevatore trova di fatto nella famiglia intervistata viene assegnata una chiave individuale secondo quanto spiegato nel Paragrafo 14.4. Ovviamente, l'acquisizione "decentralizzata" delle interviste comporta che la ditta esterna non abbia contezza del fatto che quell'individuo appena trovato in famiglia possa essere effettivamente entrato per la prima volta nel campione oppure un individuo che rientra in famiglia dopo esserne uscito temporaneamente. È quindi necessario confrontare alcune caratteristiche anagrafiche del presunto nuovo individuo (rilevate sul campo) con quelle memorizzate nel data set "storico". Le caratteristiche anagrafiche prese in considerazione sono: (1) nome, (2) cognome, (3) sesso e (4) data di nascita. Se, in base ad una certa metrica, tali caratteristiche corrispondono a quelle di un individuo già rilevato in precedenza (quindi presente nell'archivio storico), non si tratta effettivamente di un nuovo componente ma di un individuo rientrato nella famiglia, quindi gli deve essere riassegnata la chiave individuale originale (presente nell'archivio storico) e non quella generata con l'algoritmo del Paragrafo 14.4. La metrica adottata allo scopo è la seguente:

$$D\left(\mathbf{ind}^{(r)}, \mathbf{ind}^{(s)}\right) = \frac{\sum_{i=1}^{n=4} d_L\left(ind_i^{(r)}, ind_i^{(s)}\right) w_i}{\sum_{i=1}^{n=4} w_i}$$

Dove l'individuo $\mathbf{ind}^{(r)} = \left\{ind_1^{(r)} ind_2^{(r)} ind_3^{(r)} ind_4^{(r)}\right\} = \{nome\ cognome\ sesso\ data\_nascita\}$ è il presunto nuovo individuo da ricercare tra tutti quelli presenti nello storico (contrasse-

gnato con l'apice s), w_i rappresenta il peso della componente i-esima del vettore **ind** che identifica l'individuo ($w_i \geq 1$ per $i = 1, ..., 4$) e $d_L\left(ind_i^{(r)}, ind_i^{(s)}\right)$ è la distanza di Levenshtein tra le componenti i-esime $ind_i^{(s)}$ e $ind_i^{(r)}$ dei rispettivi vettori r e s. Come noto in letteratura (Dunn, 1946; Fellegi e Sunter, 1969), tale distanza tra due stringhe x e y calcola il numero minimo di modifiche elementari che consentono di trasformare la x nella y (per modifica elementare si intende la cancellazione di un carattere, la sostituzione di un carattere con un altro oppure l'inserimento di un carattere). Quindi, si calcolano le distanze tra il presunto individuo nuovo entrato in famiglia e tutti gli individui presenti nell'elenco storico. L'individuo per cui si ottiene un valore della distanza $D\left(\mathbf{ind}^{(r)}, \mathbf{ind}^{(s)}\right)$ minore di una soglia δ prestabilita è da considerarsi un individuo rientrato in famiglia. Alcune simulazioni condotte sui dati degli anni precedenti hanno portato a stabilire come soddisfacente un valore pari a $\delta = 0.1$ della soglia.

14.6 Conclusioni

La fase di importazione dei dati e di controllo, correzione e generazione delle chiavi familiari e individuali nei data set di EU-SILC è di fondamentale importanza per il successivo processo di controllo e correzione dei dati. Infatti, la quasi totalità del software sviluppato nelle fasi di lavorazione successive si basa sui data set costruiti in questa prima fase di lavorazione. L'aggiunta di alcune variabili utili al processo di produzione nonché il recupero di informazioni completamente assenti, ad esempio, per mancato contatto della famiglia (registri d'ufficio) rendono tale fase cruciale. La procedura illustrata ha carattere generale e, ad ogni nuova occasione di indagine, richiede solo un aggiornamento dei metadati dei vari modelli di rilevazione, a cura del personale tecnico dell'indagine EU-SILC, senza alcuna modifica del codice software qui brevemente descritto. Inoltre, la gestione delle interviste da parte della ditta esterna si basa su un elenco di famiglie ed individui forniti, ogni anno, dai responsabili dell'indagine prima che inizi la rilevazione. La ditta esterna distribuisce i carichi di lavoro all'interno della propria rete di rilevazione: ad ogni rilevatore viene assegnata una quota di famiglie da intervistare. Questa "decentralizzazione" non consente di gestire già durante la fase di rilevazione la generazione degli identificativi famigliari nei casi delle famiglie *split* e *fusion*, così come degli identificativi individuali per i nuovi componenti familiari, per cui è necessario anche verificare se si tratta di rientri in famiglia o meno. È dunque più che mai necessario disporre di strumenti di verifica della correttezza delle chiavi identificative basati sugli algoritmi qui presentati, fin dalla prima fase di acquisizione dei dati forniti dalla ditta esterna.

15. L'IMPATTO DELLA RILEVAZIONE PROATTIVA DELLE INFORMAZIONI SULLA COMPONENTE LONGITUDINALE[1]

15.1 Questionario elettronico e uso delle domande a conferma nelle indagini panel

Nelle indagini panel vengono raccolte le medesime informazioni, relativamente alle stesse unità di rilevazione, in occasioni di indagini successive. Nel caso di EU-SILC, per quattro anni consecutivi, famiglie e individui sono chiamati a rispondere al medesimo questionario, che rileva molte informazioni, che spaziano dalle caratteristiche dell'abitazione principale, al percorso di istruzione e occupazionale, alle fonti di reddito individuali e familiari. Nelle indagini panel, quindi, la medesima informazione viene rilevata più volte anche qualora non sussistano le condizioni per un cambiamento: si pensi alle caratteristiche dell'abitazione per una famiglia che non ha cambiato casa tra una intervista e la successiva, o al titolo di studio più alto conseguito da chi ha già terminato il proprio percorso formativo.

Tra gli obbiettivi delle indagini panel c'è la rilevazione dei cambiamenti, per descrivere i fenomeni di interesse da un punto di vista dinamico. È importante quindi che i cambiamenti che vengono rilevati siano oltre che plausibili anche rappresentativi delle effettive esperienze della popolazione. Tuttavia, per via dei diversi errori non campionari che possono avere luogo in ciascuna occasione di indagine, attribuibili ad esempio ad errori di memoria del rispondente o a errori di codifica da parte del rilevatore, è possibile che parte dei cambiamenti rilevati non siano reali ma dovuti a tali errori.

Inoltre, dal punto di vista del rispondente, la partecipazione ad una rilevazione panel è decisamente onerosa, proprio perché l'impegno richiesto è ripetuto nel tempo.

Al fine sia di migliorare la qualità delle informazioni rilevate, riducendo la rilevazione di false transizioni, sia di ridurre il carico sul rispondente, si può ricorrere già in fase di rilevazione all'uso delle informazioni rilevate in occasioni di indagine precedenti, tramite l'introduzione di domande a conferma nel questionario elettronico, il così detto *dependent interviewing* (Jackle, 2008).

Esistono diverse tipologie di domande a conferma, a seconda di come vengono sfruttate le informazioni rilevate in precedenza. Si distingue di solito tra due approcci principali: l'approccio "proattivo", per cui si chiede esplicitamente al rispondente di confermare quanto rilevato nella precedente intervista, e l'approccio "reattivo", in cui solo qualora si rilevasse una informazione incoerente con quella rilevata nella intervista precedente, si attiva una domanda di verifica per stabilire quale delle due informazioni sia da considerarsi corretta.

La prima strategia presenta il vantaggio di ridurre, almeno in parte, il carico sul rispondente, cui viene chiesto, ad esempio, di confermare di svolgere ancora la professione rilevata nell'anno precedente, anziché fornire di nuovo la descrizione della professione che poi deve essere correttamente recepita e codificata dal rilevatore. In questo esempio, anche il rilevatore sarà facilitato nel condurre la rilevazione, poiché tutte le informazioni che vengono confermate sono già codificate (Lynn et al., 2006). Il rischio di questo approccio è che il

1 I paragrafi 15.1 e 15.2 sono stati redatti da Lucia Coppola; i paragrafi 15.3 e 15.3.1 sono stati redatti da Daniela Lo Castro; i paragrafi 15.3.2, 15.4 e l'Appendice sono stati redatti da Mattia Spaziani.

rispondente, per semplicità o per non entrare in contraddizione con quanto già dichiarato in precedenza, preferisca confermare anziché rettificare l'informazione e quindi ne risulti una sottostima dei cambiamenti (Hoogendoorn, 2004). Per non incorrere in questo rischio, si può ricorrere all'approccio reattivo, per cui solo laddove l'informazione sia incompatibile con quella rilevata nella indagine precedente, se ne chiede una conferma. Ad esempio, se un intervistato dichiara un titolo di studio inferiore a quello rilevato nell'anno precedente, si può attivare una domanda di verifica per stabilire quale delle due informazioni sia da considerarsi corretta. Questo approccio però non riduce in alcun modo il carico sul rispondente o sul rilevatore, ma addirittura lo può aumentare introducendo delle domande di verifica aggiuntive. Inoltre, tale metodo può ridurre la rilevazione di transizioni impossibili ma non quella di falsi cambiamenti possibili. Entrambi gli approcci presentano quindi vantaggi e svantaggi.

Nel questionario EU-SILC si è deciso di ricorrere solo all'approccio proattivo, in quanto si è preferito sfruttare la possibilità di verificare tramite conferma le informazioni fornite nell'intervista precedente, rendendo così più agevole la somministrazione del questionario sia per il rilevatore che per il rispondente. Come atteso, l'introduzione delle domande a conferma ha avuto un impatto sulla rilevazione dei cambiamenti relativi alle variabili interessate.

In questo capitolo si descrivono le diverse strategie di conferma implementate (paragrafo 15.2), per sfruttare in modo diversificato le informazioni già rilevate a seconda del tipo di variabile presa in considerazione (paragrafo 15.3), e si discutono gli effetti delle domande a conferma sui livelli di cambiamento rilevato (paragrafo 15.4).

15.2 Questionario elettronico, domande a conferma e strategie di utilizzo delle informazioni rilevate nelle interviste precedenti

L'introduzione del questionario elettronico nella rilevazione EU-SILC, condotta in CAPI dal 2011, e con tecnica mista CATI/CAPI dal 2015, ha reso possibile l'utilizzo dei quesiti a conferma. Tra il 2012 e il 2015, periodo in cui il disegno del questionario era finalizzato a una rilevazione di tipo CAPI, le domande a conferma hanno riguardato un numero ristretto di informazioni (nome, cognome, sesso, data di nascita, comune o stato estero di nascita, prima e seconda cittadinanza, anno di arrivo in Italia e relazione di parentela), con l'obbiettivo principale di migliorare la qualità delle informazioni utilizzate per la corretta identificazione dei componenti familiari e di conseguenza migliorare la loro tracciabilità durante il periodo di osservazione del panel.

Nel 2016 è stata effettuata una rilevante revisione del questionario, con lo scopo di renderlo più adatto alla somministrazione telefonica, che a partire da quest'anno di rilevazione coinvolge un numero molto più cospicuo di famiglie (circa il 56 per cento nel 2016). Si è fatto, quindi, un uso più estensivo delle domande a conferma, che a partire dall'edizione di indagine 2016 interessano anche lo stato civile, il titolo di studio, le caratteristiche del lavoro attuale e dell'ultimo lavoro svolto, l'età di inizio del percorso lavorativo, oltre a diverse caratteristiche dell'abitazione e dell'eventuale mutuo.

A seconda del tipo di variabile per cui viene chiesta la conferma, sono state adottate diverse strategie per l'utilizzo delle informazioni rilevate nelle interviste precedenti:

1. Strategia 1 - conferma diretta: se l'informazione relativa a una specifica variabile viene confermata, la variabile viene automaticamente valorizzata con lo stesso valore rilevato nell'anno precedente, altrimenti viene somministrata di nuovo la domanda e di conseguenza compilata la variabile con il valore corretto;
2. Strategia 2 - conferma di una domanda "filtro": se l'informazione relativa a una variabile "filtro" viene confermata, si attribuisce lo stesso valore rilevato precedentemente non solo alla variabile filtro, ma anche a una serie di variabili ad essa collegate;
3. Strategia 3 - imputazione delle informazioni per cui è stato rilevato il medesimo valore per due anni: se per una variabile è stato rilevato lo stesso valore per due anni, non viene più rilevata, ma direttamente imputata dal software del questionario con il valore già rilevato in precedenza, sempre che non sussistano le condizioni per cui possa essere avvenuto un cambiamento.

La prima strategia persegue l'obbiettivo di migliorare la qualità del contenuto informativo della variabile e ha uno scarso effetto sul carico sul rispondente, che si esplica, in caso di conferma, solo tramite una più veloce compilazione del questionario elettronico da parte del rilevatore, poiché la variabile viene automaticamente valorizzata dal software. Ad esempio, se il rispondente conferma la relazione di parentela, il rilevatore deve semplicemente valorizzare la domanda a conferma (che presenta due modalità "sì" o "no") e non è necessario che proceda con la codifica della variabile "relazione di parentela" che presenta 20 modalità.

La seconda strategia, oltre a verificare e validare l'informazione rilevata tramite la domanda filtro, riduce significativamente il carico sul rispondente, poiché alcuni quesiti non vengono somministrati all'intervistato. È il caso dello stato civile. Qualora il rispondente confermi lo stato civile rilevato nell'anno precedente, non viene chiesto nuovamente l'anno e il mese di conseguimento dello stato civile, lo stato civile precedente quello attuale (per coniugati, separati, divorziati o vedovi), il motivo della separazione per i separati di fatto e il regime patrimoniale per i coniugati. Ad esempio, ai coniugati che confermano di avere lo stesso stato civile dell'anno precedente, che rappresentano il 56 per cento degli individui di almeno 16 anni intervistati sia nel 2015 che nel 2016, non vengono somministrati 4 quesiti (ovvero il mese e anno del matrimonio, lo stato civile precedente il matrimonio e il regime patrimoniale). La medesima strategia è stata usata anche per le famiglie che confermano di sostenere la rata di un mutuo, che rappresentano l'11,4 per cento delle famiglie intervistate nel 2015 e nel 2016, cui non si chiede nuovamente l'ammontare del mutuo, l'anno di stipula e la durata. Analogamente a chi conferma il titolo di studio conseguito (circa il 95 per cento degli individui di almeno 16 anni intervistati nel 2015 e nel 2016) non vengono somministrate alcune domande relative all'istruzione. In tutti questi casi, quindi, si assume che qualora venga confermata la variabile filtro, tutte le altre ad essa associate siano state rilevate correttamente nell'anno precedente e non sia necessario rilevarle nuovamente.

La terza strategia si applica al solo sottoinsieme di famiglie e individui che si trovano alla terza o quarta intervista, e interessa le caratteristiche dell'abitazione e dell'ultima occupazione svolta (per chi non è occupato al momento dell'intervista). Per le famiglie che dichiarano di non aver cambiato l'abitazione e di non aver effettuato lavori strutturali, non vengono più rilevate le variabili relative alle caratteristiche dell'abitazione per cui è stato rilevato lo stesso valore per due anni consecutivi (ad esempio il tipo di abitazione, l'anno di costruzione dell'immobile, il numero di stanze ecc. per un totale di 14 variabili). La percentuale di famiglie interessate da questa strategia varia, di variabile in variabile, secondo il numero di casi per cui è stato rilevato lo stesso valore per due anni consecutivi. Analoga-

mente, agli individui per cui per due volte è stata rilevata la medesima occupazione passata, non viene somministrata l'intera sezione del questionario sull'ultima occupazione (per un totale di 13 quesiti). Chiaramente l'utilizzo di questo tipo di strategia è ristretto solo a quelle caratteristiche individuali o familiari per cui si può stabilire che sotto determinate condizioni non siano avvenuti cambiamenti tra un'intervista e la successiva.

Le diverse strategie messe in atto hanno sicuramente un impatto sulla qualità del dato e, in ottica strettamente longitudinale, sulla stima delle transizioni e dei cambiamenti nel periodo di osservazione.

15.3 Effetti delle domande a conferma sulle transizioni rilevate tra coppie di anni

Per valutare la confrontabilità della rilevazione 2016 (anno in cui il questionario è stato ridisegnato per la somministrazione CAPI/CATI) con quella degli anni precedenti, in primo luogo si confrontano le domande a conferma somministrate in modo analogo sia nel questionario 2016, sia in quello implementato negli anni precedenti (Tavola 15.1). La percentuale di famiglie teoricamente già intervistate l'anno precedente (o due anni prima nel caso in cui l'anno precedente non sia stato possibile effettuare l'intervista) che non confermano di essere state effettivamente intervistate è piuttosto esigua, inferiore al 2 per cento negli anni 2013-2015, raggiunge il 2,3 per cento nel 2016, interessando meno di 300 famiglie.

Per quanto riguarda le variabili demografiche, richieste a conferma già a partire dal 2012, si osserva che la percentuale di individui che non confermano il dato rilevato nell'anno precedente riguarda poche unità, in tutti gli anni di rilevazione presi in esame. Il questionario 2016 presenta risultati del tutto analoghi a quelli delle rilevazioni precedenti. Le mancate conferme più frequenti si osservano relativamente alla relazione di parentela, per cui in meno del 2 per cento dei casi si osserva una rettifica, dovuta probabilmente a un errore di codifica che è facilmente ammissibile considerando che la variabile è composta al momento da 20 modalità.

Questi dati mostrano che, laddove confrontabili, le conferme della rilevazione del 2016 sono analoghe a quelle degli anni precedenti. Pertanto, si può procedere con un'analisi dell'impatto dei nuovi quesiti a conferma, introdotti nel 2016, sulle coerenze longitudinali

Tavola 15.1 - Distribuzione di alcune domande a conferma per anno di rilevazione

	2013		2014		2015		2016	
	N	%	N	%	N	%	N	%
INTERVISTA PRECEDENTE								
Sì	12.117	98,1	13.146	98,3	12.613	99,1	12.263	97,7
No	232	1,9	225	1,7	115	0,9	287	2,3
RELAZIONE DI PARENTELA								
Sì	29.108	98,4	31.213	98,4	29.905	98,1	28.321	98,8
No	468	1,6	514	1,6	571	1,9	336	1,2
COMUNE DI NASCITA								
Sì	27.787	99,3	29.865	99,6	29.266	99,9	1.1011	99,6
No	188	0,7	110	0,4	27	0,1	44	0,4
STATO ESTERO DI NASCITA								
Sì	1.704	98,8	1.891	99,5	1.920	99,0	848	98,5
No	21	1,2	10	0,5	20	1,0	13	1,5
PRIMA CITTADINANZA								
Sì	1.197	98,8	1.404	99,0	1.258	98,8	1.180	99,4
No	15	1,2	14	1,0	15	1,2	7	0,6
SECONDA CITTADINANZA								
Sì	307	99,0	361	99,5	376	99,5	363	98,4
No	3	1,0	2	0,6	2	0,5	6	1,6

delle variabili interessate. A tal fine, si mettono a confronto le coerenze rilevate sulle variabili "grezze" (ovvero così come sono rilevate, prima di subire l'usuale processo di correzione e trattamento dati) nelle coppie di anni che precedono l'uso estensivo delle domande a conferma (2013-2014, 2014-2015) e in quella interessata dal cambiamento del questionario (2015-2016). Per il 2016, inoltre, viene mostrata la percentuale di conferme ottenute durante le interviste.

Nelle tavole di seguito si mostrano le coerenze osservate solo sul sottoinsieme di famiglie o individui che è stato effettivamente intervistato in entrambi gli anni, i così detti compresenti. Per omogeneità, anche le distribuzioni delle domande a conferma vengono mostrate relativamente al solo sottoinsieme di famiglie o individui compresenti tra due anni contigui, anche se in realtà vengono somministrate anche alle famiglie per cui non è stato possibile condurre l'intervista nell'anno precedente, ma sono state intervistate due anni prima.

Di seguito, verrà commentata solo una selezione di variabili a conferma, sia a livello familiare, sia a livello individuale. Le tavole relative a tutte le variabili interessate dalle domande a conferma sono disponibili nell'Appendice al capitolo.

15.3.1 Questionario familiare

Per quanto riguarda le variabili familiari, si procede con l'implementazione delle diverse strategie di domande a conferma, solo se la famiglia dichiara di non aver cambiato l'abitazione principale e di non aver effettuato modifiche strutturali dell'abitazione. Nel questionario familiare si ricorre alla conferma diretta di 5 variabili (strategia 1), alla conferma della sola variabile filtro in un caso (strategia 2) e dell'imputazione di 14 caratteristiche dell'abitazione nei casi in cui sia già stato rilevato lo stesso valore per due anni (strategia 3).

Tra le variabili interessate solo dalla conferma diretta, si commentano solo le due variabili che riguardano tutte le famiglie[2], ovvero il titolo di godimento dell'abitazione e le spese condominiali (Tavola 15.2).

Con riferimento al titolo di godimento dell'abitazione, si osserva nel 2016 che solo l'1,4 per cento delle famiglie non conferma quanto rilevato nel 2015. La mancata conferma può essere attribuibile o a un effettivo cambiamento del titolo di godimento dell'abitazione (ad esempio un affittuario che diventa proprietario della medesima abitazione) oppure ad un errore di rilevazione dell'anno precedente.

Per comparabilità con le coppie di anni precedenti, non interessate dall'uso del quesito a conferma, si mettono a confronto tutte le famiglie intervistate tra il 2015 e il 2016, incluse quelle cui non viene somministrato il quesito a conferma poiché hanno cambiato abitazione. Per queste famiglie si osserva un cambiamento del titolo di godimento dell'abitazione nel 2 per cento dei casi, attribuibile per lo più a una mancata conferma dell'informazione rilevata nell'anno precedente (142 casi[3]) e in parte ad un cambiamento dell'abitazione cui è corrisposto anche un cambiamento del titolo di godimento (90 casi). Nelle coppie di anni precedenti l'introduzione del quesito a conferma, invece, si osservavano valori incoerenti, e quindi possibili transizioni, nel 7-8 per cento dei casi.

2 Le altre tre variabili riguardano solo il sottoinsieme di famiglie che vivono in affitto, usufrutto o uso gratuito e sono presentate nelle tavole in Appendice.

3 Le mancate conferme effettivamente rilevate sono 155, ma 13 famiglie, pur non confermando il dato rilevato nell'anno precedente, forniscono la stessa risposta. Tali casi sono probabilmente attribuibili ad errori di compilazione da parte del rilevatore.

Con riferimento alle spese condominiali, si osserva una percentuale di mancate conferme più elevata, pari al 4,6 per cento delle famiglie intervistate nel 2015 e nel 2016. La percentuale di incoerenze osservate tra il 2015 e il 2016 si attesta al 5,2 per cento dei casi, attribuibili principalmente alla mancata conferma dell'informazione rilevata nell'anno precedente (508 su 599), di poco inferiore ai livelli osservati nelle coppie di anni precedenti (6-7 per cento).

Si nota quindi che i livelli di mancata conferma variano sensibilmente a seconda della variabile, così come il loro impatto sulle differenze rilevate tra coppie di anni.

Tavola 15.2 - Esempio di variabili familiari a conferma diretta (strategia 1), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
TITOLO DI GODIMENTO ABITAZIONE								
Sì	10.965	98,6	11.265	98,0	11.997	92,7	11.600	92,1
No	155	1,4	232	2,0	942	7,3	1002	8,0
SOSTIENE SPESE CONDOMINIALI								
Sì	10.612	95,4	10.898	94,8	12.101	93,5	11.703	92,9
No	508	4,6	599	5,2	838	6,5	899	7,1

La conferma della variabile filtro viene utilizzata nel questionario familiare solo relativamente al pagamento della rata del mutuo o prestito per l'acquisto o ristrutturazione dell'abitazione e riguarda un sottoinsieme esiguo della popolazione, ovvero solo le famiglie che nell'anno precedente hanno dichiarato di pagare il mutuo (Tavola 15.3). La percentuale di famiglie che non conferma di sostenere una rata di mutuo è relativamente alta, circa il 15 per cento. Anche in questo caso, la mancata conferma può essere attribuita ad un errore di rilevazione nell'anno precedente o a un effettivo cambiamento, ad esempio per le famiglie che hanno estinto il proprio mutuo/prestito. Alle famiglie che confermano di sostenere ancora un mutuo o prestito, non viene più chiesto l'anno di stipula, l'ammontare o la durata, che vengono invece imputati uguali a quelli rilevati nell'anno precedente.

Considerando tutte le famiglie che risultano sostenere un mutuo o prestito sia nel 2015 che nel 2016 (quindi non solo quelle cui è stato somministrato il quesito a conferma), solo una percentuale esigua (circa il 2-3 per cento) mostra valori diversi relativamente all'anno di stipula, alla durata o all'ammontare del mutuo o prestito. Le differenze tra queste variabili rilevate nelle coppie di anni precedenti mostrano livelli di incoerenza decisamente elevati, attribuibili a diversi errori di tipo non campionario, quali errori di digitazione da parte dell'intervistatore o di memoria da parte del rispondente. Si consideri, ad esempio, che il rispondente alle domande del questionario familiare potrebbe cambiare da un anno all'altro e che questo tipo di informazioni richiedono un notevole sforzo mnemonico e sono soggette ad approssimazioni.

L'introduzione del quesito a conferma sulla sola domanda filtro ha quindi un impatto notevole sulle coerenze longitudinali relative alle caratteristiche del mutuo. Ricorrendo alla verifica tramite conferma della sola domanda filtro, si perde l'opportunità di migliorare la qualità delle informazioni relative alle altre caratteristiche del mutuo, che si è visto presentare delle difficoltà di rilevazione. Tuttavia si è preferito accettare un margine di errore per ridurre notevolmente il carico su questo sottoinsieme esiguo di rispondenti, evitando di sottoporre nuovamente quesiti che possono richiedere un notevole sforzo mnemonico.

Tavola 15.3 - Variabili familiari con conferma di domanda filtro (strategia 2), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
Domanda filtro								
CONFERMA DI SOSTENERE RATE MUTUO/PRESTITO								
Sì	1.311	84,7						
No	236	15,3						
ANNO INIZIO MUTUO								
Sì			1.348	97,8	1.392	78,4	1.154	71,0
No			31	2,3	383	21,6	471	29,0
CAPITALE INIZIALE MUTUO								
Sì			1.260	91,4	1.091	61,5	890	54,8
No			119	8,6	684	38,5	735	45,2
DURATA MUTUO								
Sì			1.341	97,2	1.448	81,6	1.270	78,2
No			38	2,8	327	18,4	355	21,9

Per l'applicazione della terza strategia è necessario suddividere in due gruppi il collettivo delle famiglie che non hanno cambiato abitazione: (i) le famiglie per cui l'informazione sia stata rilevata una sola volta o non sia stata confermata in interviste successive, per cui si procede con la somministrazione del quesito a conferma diretta (ovvero strategia 1); (ii) le famiglie per cui la caratteristica dell'abitazione è stata già rilevata due volte, per cui l'informazione di interesse non viene richiesta ma automaticamente imputata uguale a quella già rilevata in precedenza. Gli effetti della strategia 3 vanno quindi letti insieme a quelli attribuibili alla strategia 1, in quanto viene utilizzata l'una o l'altra a seconda della casistica (Tavola 15.4).

Si nota che la combinazione di queste strategie ha un impatto notevole sulle coerenze rilevate tra due anni contigui. In particolare, si pensi alla tipologia di abitazione (villa, appartamento, ecc.) per cui si rilevavano incoerenze superiori al 20 per cento a fronte del 3 per cento dopo l'introduzione della domanda a conferma, secondo la strategia 1 o 3 a seconda della casistica. Dei 370 cambiamenti rilevati tra il 2015 e 2016, meno della metà (157) sono attribuibili a una reale mancata conferma dell'informazione rilevata nell'anno precedente, mentre i rimanenti casi sono attribuibili a cambiamenti di abitazione. Le altre variabili presentate nella tavola a titolo esemplificativo subiscono un impatto ancora maggiore, poiché sono variabili ancor più soggette ad errori di memoria o approssimazioni da parte del rispondente o di compilazione da parte del rilevatore.

Tavola 15.4 - Esempio di variabili familiari con imputazione delle variabili già rilevate due volte (strategia 1 e 3), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
TIPO DI ABITAZIONE								
Sì	5.014	96,6	11.127	96,8	10.347	80,0	9.609	76,3
No	176	3,4	370	3,2	2.592	20,0	2.993	23,8
ANNO DA CUI VIVE NELL'ABITAZIONE								
Sì	5.785	97,6	10.979	95,5	8.834	68,3	7.960	63,2
No	143	2,4	518	4,5	4.105	31,7	4.642	36,8
ANNO DI COSTRUZIONE DELL'EDIFICIO								
Sì	5.541	97,5	10.701	93,1	8.858	68,5	7.883	62,6
No	145	2,6	796	6,9	4.081	31,5	4.719	37,5
NUMERO DI STANZE								
Sì	5.832	97,7	11.120	96,7	9.030	69,8	8.170	64,8
No	140	2,3	377	3,3	3.909	30,2	4.432	35,2
METRI QUADRATI DELL'ABITAZIONE								
Sì	7.116	97,4	11.004	95,7	6.685	51,7	5.162	41,0
No	192	2,6	492	4,3	6254	48,3	7.440	59,0

15.3.2 Questionario individuale

Nel questionario individuale si fa ricorso alla conferma diretta per 14 variabili (strategia 1), alla conferma della sola variabile filtro per 5 variabili (strategia 2) e all'imputazione diretta delle variabili già confermate in precedenza per 3 variabili, oltre che per l'intera sezione sulle caratteristiche dell'ultimo lavoro svolto (strategia 3).

Come per le variabili familiari, anche a livello individuale si commentano solo alcune variabili, rimandando all'Appendice per un quadro esaustivo. Con riferimento alle domande a conferma diretta (strategia 1, Tavola 15.5), ad esempio agli individui che dichiarano di essere occupati sia nel 2015 che nel 2016, si chiede conferma della tipologia di lavoro dichiarata nel 2015. Solo un 2 per cento dei rispondenti non conferma. Questi casi potrebbero essere dovuti non solo a un'errata codifica dell'occupazione nell'anno precedente, ma anche ad un effettivo cambiamento dell'attività lavorativa. Si nota che i cambiamenti dell'attività lavorativa principale, tra chi rimane occupato nei due anni, scendono da circa il 10 per cento osservato nelle coppie di anni precedenti l'introduzione della domanda a conferma, al 2 per cento. Le caratteristiche del lavoro che mostrano una maggiore riduzione dei cambiamenti sono quelle che presentano maggiori difficoltà di codifica, ovvero la professione e il settore di attività (i cambiamenti passano da circa il 30-40 per cento degli anni precedenti a 4-5 per cento osservato tra il 2015 e il 2016).

Per quanto riguarda l'uso delle domande filtro (strategia 2), si mostra, a titolo esemplificativo, l'effetto dell'introduzione della conferma sullo stato civile. La conferma su questa variabile non solo ha l'effetto di ridurre i cambiamenti di stato civile da circa il 4 per cento a poco più dell'1 per cento, ma rende quasi trascurabili i cambiamenti delle altre variabili che vengono automaticamente imputate uguali a quelle rilevate nell'anno precedente se lo stato civile viene confermato (Tavola 15.6). Da una parte, nel caso in cui lo stato civile non sia cambiato, è corretto assumere che anche le altre caratteristiche siano rimaste identiche. Tuttavia, per quanto riguarda variabili per cui è più facile che vengano commessi errori di memoria da parte del rispondente o di compilazione del questionario da parte del rilevatore, come ad esempio la data di matrimonio, si perde l'occasione di verificare l'informazione, assumendo che quella già rilevata sia corretta.

Tavola 15.5 - Esempio di variabili individuali a conferma diretta (strategia 1), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
TIPO DI LAVORO								
Sì	9.008	97,9	9.016	98,0	9.871	90,8	9.087	88,9
No	191	2,1	184	2,0	1.001	9,2	1.134	11,1
LAVORO TEMPO PIENO/PARZIALE								
Sì	8.910	96,9	8.985	97,7	10.172	93,6	9.508	93,0
No	289	3,1	213	2,3	700	6,4	713	7,0
PROFESSIONE								
Sì	8.708	94,7	8.733	95,0	7.188	66,1	5.969	58,6
No	488	5,3	464	5,1	3.680	33,9	4.226	41,5
SETTORE PUBBLICO/PRIVATO								
Sì	6.971	98,6	9.085	98,8	10.516	96,7	9.817	96,1
No	102	1,4	115	1,3	356	3,3	404	4,0
NACE								
Sì	8.791	96,0	8.814	96,2	7.087	65,6	5.773	56,8
No	367	4,0	345	3,8	3.725	34,5	4.392	43,2

Tavola 15.6 - Esempio di variabili individuali a conferma di domanda filtro (strategia 2), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
Domanda filtro:								
CONFERMA STATO CIVILE								
Sì	22.111	98,1						
No	427	1,9						
STATO CIVILE								
Sì			22.238	98,7	24.847	95,8	24.733	96,4
No			300	1,3	1.097	4,2	913	3,6
MOTIVO SEPARAZIONE (PER I SEPARATI DI FATTO)								
Sì			317	98,8	165	79,7	145	68,4
No			4	1,3	42	20,3	67	31,6
ANNO DI STATO CIVILE								
Sì			15.702	99,4	15.480	86,7	15.018	85,7
No			91	0,6	2366	13,3	2.509	14,3
MESE DI STATO CIVILE								
Sì			15.642	99,5	15.243	85,9	14.671	84,2
No			76	0,5	2.504	14,1	2750	15,8
REGIME PATRIMONIALE (PER I CONIUGATI)								
Sì			14.145	99,7	14.603	90,9	14.257	90,9
No			41	0,3	1.458	9,1	1.434	9,1
STATO CIVILE PRECEDENTE IL MATRIMONIO (PER I CONIUGATI)								
Sì			12.675	100,0	14.333	99,3	14.069	99,4
No			3	0,0	98	0,7	89	0,6

Un'ultima casistica che si vuole mostrare anche in riferimento al questionario individuale è quella relativa all'imputazione di variabili già rilevate e confermate in precedenti interviste. A coloro che dichiarano di non essere occupati al momento dell'intervista ma hanno svolto un lavoro in passato, si richiedono le caratteristiche dell'ultimo lavoro svolto. Qualora tali informazioni siano già state rilevate si procede in due modi, a seconda del numero di interviste somministrate all'individuo: (1) se le informazioni sono già state rilevate una volta e l'individuo dichiara di non aver svolto nessuna attività dall'ultima intervista, viene chiesta conferma della professione dichiarata nell'intervista precedente e, se confermata, viene imputato il resto della sezione con le informazioni rilevate precedentemente (strategia 2, che interessa il 21,4 per cento degli individui di almeno 16 anni intervistati nel 2015 e 2016); (2) se l'individuo è almeno alla terza intervista, dichiara di non aver svolto alcuna attività dall'ultima intervista e all'intervista precedente ha già dato conferma della professione svolta in passato, allora si imputa direttamente l'intera sezione con le informazioni già a disposizione (strategia 3, che interessa l'11,9 per cento degli individui di almeno 16 anni intervistati nel 2015 e 2016). Anche in questo caso, gli effetti della combinazione delle due strategie implementate vanno valutati nel loro complesso e si nota una forte riduzione delle incoerenze rilevate tra anni contigui (Tavola 15.7).

Tavola 15.7 - Esempi di variabili individuali con imputazione delle conferme precedenti (strategia 2 e 3), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
Domanda filtro:								
CONFERMA ULTIMO LAVORO SVOLTO								
Sì	7.494	97,2						
No	213	2,8						
HAI MAI SVOLTO UN LAVORO								
Sì			11.198	96,7	12.216	93,0	12.025	91,5
No			380	3,3	918	7,0	1111	8,5
TIPO ULTIMO LAVORO								
Sì			7.804	98,8	7.251	88,2	7.193	87,1
No			93	1,2	967	11,8	1.063	12,9
PROFESSIONE ULTIMO LAVORO								
Sì			7.557	95,7	4.974	60,5	4.607	55,9
No			340	4,3	3244	39,5	3.636	44,1
NACE ULTIMO LAVORO								
Sì			7.583	96,2	4.920	60,0	4.491	54,6
No			301	3,8	3278	40,0	3.742	45,5

15.4 Conclusioni

EU-SILC è una rilevazione che raccoglie un ampio contenuto informativo, sulle medesime unità campionarie, che vengono intervistate fino a quattro volte consecutive. L'informazione rilevata in ciascun anno dovrebbe essere coerente, dal punto di vista longitudinale, con quella rilevata precedentemente. Ad esempio, per chi non ha cambiato abitazione è presumibile che le caratteristiche dell'abitazione non siano cambiate, oppure i cambiamenti dello stato civile o del titolo di studio possono seguire solo determinati percorsi (il titolo di studio più alto conseguito non può diminuire da un anno all'altro, così come chi non è più celibe non può tornare ad esserlo in occasioni di interviste successive). Tuttavia, gli errori non campionari, dall'errore di compilazione da parte del rilevatore, all'errore di memoria del rispondente, conducono alla rilevazione di cambiamenti longitudinali per cui non sempre è possibile distinguere tra un cambiamento reale o un errore di rilevazione. Inoltre, anche laddove non sia accettabile un cambiamento, ad esempio una diversa tipologia dell'abitazione principale qualora questa non sia cambiata, può diventare arbitrario stabilire quale sia l'informazione corretta tra quelle incoerenti rilevate in occasioni di indagine diverse.

L'utilizzo estensivo del questionario elettronico per l'introduzione delle domande a conferma consente non solo di ridurre il carico sul rispondente, ma anche di migliorare la qualità dell'informazione rilevata. Ad esempio, in tutti quei casi in cui non è ammissibile un cambiamento, la conferma o meno di una determinata caratteristica consente di attribuire una maggiore affidabilità all'informazione rilevata nell'occasione di indagine più recente e quindi alle correzioni longitudinali retrospettive (si veda capitolo 16 per un approfondimento).

L'implementazione delle diverse strategie di domande a conferma in occasione della rilevazione 2016 mostra che queste hanno avuto un impatto rilevante sulle coerenze longitudinali relative alle diverse variabili interessate. In particolare, i risultati che emergono a livello sia di variabili familiari, sia di variabili individuali mostrano che: (i) i tassi di mancata conferma sono contenuti, per lo più inferiori al 5 per cento; (ii) i cambiamenti tra anni contigui rilevati su variabili sottoposte a conferma è sensibilmente inferiore rispetto a quelli os-

servati negli anni che precedono la ristrutturazione del questionario, con effetti più evidenti in corrispondenza delle variabili che presentano maggiori difficoltà di rilevazione, perché richiedono uno sforzo mnemonico al rispondente, o perché soggette ad approssimazioni (ad esempio le variabili quantitative) o perché di difficile codifica da parte del rilevatore (come ad esempio la professione e l'attività economica).

La preoccupazione maggiore nell'introdurre i quesiti a conferma è quella del possibile condizionamento del rispondente, ovvero si corre il rischio che il rispondente confermi quanto già rilevato in occasioni precedenti, per brevità o per non entrare in contraddizione con quanto dichiarato in precedenza. Tuttavia, si nota come diversi quesiti a conferma, pur sottostando alla stessa logica di somministrazione, presentano effetti sulle incoerenze rilevate piuttosto variabili. Quindi l'effetto condizionamento dei quesiti a conferma non è sistematico e si può assumere che sia contenuto. D'altra parte, le domande a conferma rappresentano un'opportunità per verificare e migliorare la qualità delle informazioni rilevate, e forniscono uno strumento per individuare le false transizioni e stabilire delle strategie di correzione più precise (si veda capitolo 16).

APPENDICE

Tavola 15.1A - Altre variabili familiari a conferma diretta (strategia 1), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
PROPRIETARIO DELL'ABITAZIONE DI FAMIGLIA IN USUFRUTTO/TITOLO GRATUITO								
Sì	1.252	98,9	1.262	98,9	1.122	96,0	1.079	95,6
No	14	1,1	14	1,1	47	4,0	50	4,4
PROPRIETARIO DELL'ABITAZIONE DI FAMIGLIA IN AFFITTO								
Sì	1.410	97,8	1.440	97,8	1.676	92,7	1.592	89,2
No	32	2,2	32	2,2	133	7,4	193	10,8
TIPOLOGIA CONTRATTO DI AFFITTO								
Sì	1.368	98,1	1.389	98,0	1.393	80,8	1.325	78,9
No	27	1,9	29	2,1	331	19,2	355	21,1

Tavola 15.2A - Altre variabili familiari con imputazione delle variabili già rilevate due volte (strategia 1 e 3), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
CUCINA ABITABILE								
Sì	7.044	98,4	11.313	98,4	11.945	92,3	11.496	91,2
No	113	1,6	184	1,6	994	7,7	1.106	8,8
GABINETTO INTERNO								
Sì	7.118	99,5	11.459	99,7	12.824	99,1	12.473	99,0
No	39	0,5	38	0,3	115	0,9	129	1,0
VASCA DA BAGNO O DOCCIA								
Sì	7.136	99,7	11.472	99,8	12.865	99,4	12.498	99,2
No	21	0,3	25	0,2	74	0,6	104	0,8
DUE O PIU' BAGNI								
Sì	6.872	96,0	11.158	97,1	11.440	88,4	10.785	85,6
No	285	4,0	339	3,0	1.499	11,6	1817	14,4
CANTINA SOLAIO O SOFFITTA								
Sì	6.907	96,5	11.162	97,1	10.846	83,8	10.208	81,0
No	250	3,5	335	2,9	2.093	16,2	2394	19,0
TERRAZZA O BALCONE								
Sì	7.005	97,9	11.284	98,2	11.684	90,3	11.184	88,8
No	152	2,1	213	1,9	1.255	9,7	1418	11,3
GIARDINO PRIVATO								
Sì	6.874	96,1	11.125	96,8	11.010	85,1	10.370	82,3
No	283	4,0	372	3,2	1929	14,9	2232	17,7
ACQUA CALDA								
Sì	7.119	99,5	11.446	99,6	12.706	98,2	12.351	98,0
No	38	0,5	51	0,4	233	1,8	251	2,0
GARAGE PRIVATO O POSTO AUTO								
Sì	6.894	96,3	11.159	97,1	11.218	86,7	10.608	84,2
No	263	3,7	338	2,9	1.721	13,3	1.994	15,8

Tavola 15.3A - Altre variabili individuali a conferma diretta (strategia 1), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
POSIZIONE (DIPENDENTE)								
Sì	6.693	97,6	6.723	98,0	6.886	88,2	6.290	86,1
No	165	2,4	135	2,0	922	11,8	1.015	13,9
COORDINAMENTO (DIPENDENTE)								
Sì	6.043	88,1	6.043	88,1	6.958	89,1	6.282	86,0
No	815	11,9	815	11,9	850	10,9	1.023	14,0
CONTRATTO/ACCORDO VERBALE (DIPENDENTE)								
Sì	6.829	99,6	6.844	99,8	7.756	99,3	7.232	99,0
No	29	0,4	13	0,2	52	0,7	73	1,0
CONTRATTO TEMPO DET/IND (DIPENDENTE)								
Sì	6.684	97,5	6.697	97,7	7.328	93,9	6.823	93,4
No	174	2,5	161	2,4	480	6,2	482	6,6
HA DIPENDENTI (AUTONOMO)								
Sì	1.735	91,5	1.736	91,5	1.925	90,9	1.743	88,3
No	162	8,5	162	8,5	192	9,1	230	11,7
NUMERO ADDETTI (CLASSI)								
Sì	8.858	97,0	8.914	97,6	8.159	75,8	7.200	71,3
No	277	3,0	222	2,4	2.607	24,2	2.898	28,7
CITTADINANZA ITALIANA								
Sì	22.426	99,5	22.425	99,5	25.779	99,4	25.541	99,6
No	112	0,5	112	0,5	161	0,6	101	0,4
HA UNA SECONDA CITTADINANZA								
Sì	252	66,8	22.295	98,9	25.738	99,2	25.494	99,4
No	125	33,2	242	1,1	196	0,8	144	0,6
RIMASTO SEMPRE IN ITALIA								
Sì	986	99,7	986	99,7	1.783	95,4	1.655	95,3
No	3	0,3	3	0,3	86	4,6	82	4,7

La Tavola 15.4A mostra un'anomalia in corrispondenza delle conferme relative al titolo di studio, che ha luogo tramite due domande separate, a seconda che l'individuo abbia dichiarato di avere conseguito un titolo post-universitario o meno. Coloro a cui viene chiesta la conferma sul titolo post-universitario rilevato l'anno precedente presentano un livello di conferma particolarmente basso (66 individui pari al 12 per cento). L'anomalia consiste nel fatto che tra coloro che non confermano l'informazione dell'anno precedente, quasi la metà effettivamente dichiara di non aver alcun titolo post-universitario (205 individui), una piccola parte rettifica la codifica del titolo (72 individui) e una parte cospicua, pur non confermando, fornisce nuovamente la stessa informazione dell'anno precedente (186 individui). Questo ultimo tipo di casistica si osserva anche per altre variabili, ma interessa generalmente pochi casi, mentre per questa variabile ha una frequenza rilevante. In pratica, se questi casi venissero considerati come conferme, la percentuale di mancate conferme scenderebbe dall'88 al 52 per cento, rimanendo comunque particolarmente elevata.

Per quanto riguarda il titolo di studio inferiore a quello post-universitario, che riguarda la maggior parte del campione, si nota che le mancate conferme sono circa il 5 per cento e i livelli di transizione scendono dal 18 per cento del 2014-2015 al 4 per cento del 2015-2016 (per il 2013-2014 non è possibile effettuare il confronto perché nel 2014 è stata ristrutturata la sezione del questionario relativa all'istruzione e quindi le variabili da rilevazione non sono direttamente confrontabili). Le altre variabili della sezione, che vengono imputate uguali all'anno precedente se il titolo di studio viene confermato, presentano livelli di cambiamento piuttosto bassi.

Tavola 15.4A - Altre variabili individuali a conferma di domanda filtro (strategia 2), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
Domanda filtro: CONFERMA TITOLO DI STUDIO POST-UNIVERSITARIO								
Sì	66	12,5						
No	463	87,5						
Domanda filtro: CONFERMA TITOLO DI STUDIO PRE-UNIVERSITARIO								
Sì	21.372	95,1						
No	1.094	4,9						
TITOLO DI STUDIO								
Sì			21.542	95,6	19.175	82,2		
No			991	4,4	4145	17,8		
TIPO DI DIPLOMA								
Sì			1.883	99,8	1.580	81,8		
No			3	0,2	351	18,2		
ALTRO TIPO DI DIPLOMA								
Sì			45	97,8	23	71,9		
No			1	2,2	9	28,1		
ANNO CONSEGUIMENTO TITOLO DI STUDIO								
Sì			17.400	96,7	11.765	63,4		
No			602	3,3	6793	36,6		
ETA' TITOLO DI STUDIO								
Sì			3.561	99,3	1.427	64,3		
No			25	0,7	791	35,7		
TITOLO DI STUDIO POST-UNIVERSITARIO								
Sì			2.808	86,8	2.887	90,0		
No			426	13,2	320	10,0		
ANNO CONSEGUIMENTO TITOLO DI STUDIO POST-UNIVESITARIO								
Sì			183	62,5	256	70,5		
No			110	37,5	107	29,5		
ETA' TITOLO DI STUDIO POST-UNIVERSITARIO								
No			1	100,0	1	100,0		
Domanda filtro: CONFERMA DI AVER/NON AVER EFFETTUATO CORSO REGIONE								
Sì	21.943	98,1						
No	419	1,9						
CORSO REGIONE								
Sì			21.944	98,1	21.519	92,8	19.698	90,7
No			419	1,9	1677	7,2	2.010	9,3
DURATA CORSO REGIONE								
Sì			1547	100,0	804	73,6	581	58,9
No					288	26,4	406	41,1
TITOLO PER CORSO REGIONE								
Sì			1548	100,0	902	82,8	688	69,6
No					187	17,2	301	30,4

Tavola 15.5A - Altre variabili individuali con imputazione delle conferme precedenti (strategia 2 e 3), percentuale di conferme e distribuzione delle coerenze tra coppie di anni

	CONFERMA		COERENZA TRA COPPIE DI ANNI					
	2016		2015-2016		2014-2015		2013-2014	
	N	%	N	%	N	%	N	%
Domanda filtro:								
CONFERMA ULTIMO LAVORO SVOLTO								
Sì	7.494	97,2						
No	213	2,8						
AVEVA DIPENDENTI ULTIMO LAVORO								
Sì			1.230	99,8	963	90,0	933	87,9
No			3	0,2	107	10,0	128	12,1
POSIZIONE ULTIMO LAVORO								
Sì			5.983	98,9	5.334	88,0	5.200	86,8
No			66	1,1	725	12,0	791	13,2
COORDINAMENTO ULTIMO LAVORO								
Sì			5.975	98,8	5.477	90,4	5.275	88,0
No			72	1,2	580	9,6	719	12,0
CONTRATTO/ACCORDO ULTIMO LAVORO								
Sì			6.027	99,7	5.899	97,1	5.779	96,1
No			19	0,3	178	2,9	237	3,9
TIPO CONTRATTO ULTIMO LAVORO								
Sì			5.991	99,2	5.651	93,0	5.594	93,0
No			50	0,8	423	7,0	422	7,0
NUMERO ADDETTI ULTIMO LAVORO (CLASSI)								
Sì			6.909	88,4	5.692	70,0	5.459	67,0
No			909	11,6	2.443	30,0	2.687	33,0
NUMERO ADDETTI ULTIMO LAVORO (DETTAGLIO)								
Sì			2.962	98,4	1.578	60,1	1.265	49,8
No			49	1,6	1.046	39,9	1.273	50,2
ETA' PRIMO LAVORO								
Sì	11.555	94,7	17.326	96,6	10.271	51,2	8.012	41,6
No	642	5,3	604	3,4	9.791	48,8	11.259	58,4
NATO IN ITALIA								
Sì	7.076	99,9	25.609	99,9	25.889	99,8	25.609	99,9
No	10	0,1	33	0,1	51	0,2	33	0,1
CITTADINO ITALIANO DALLA NASCITA								
Sì	7.384	97,2	24.142	98,5	24.345	98,6	24.142	98,5
No	211	2,8	376	1,5	359	1,5	376	1,5

16. IL TRATTAMENTO DEI DATI LONGITUDINALI[1]

16.1 Introduzione

Il trattamento del dato longitudinale pone delle sfide aggiuntive rispetto a quelle di un'indagine esclusivamente trasversale, in quanto è necessario assicurare la coerenza delle caratteristiche individuali e familiari e i relativi cambiamenti nel tempo tenendo conto simultaneamente delle coerenze tra variabili sia a livello trasversale che longitudinale.

Il controllo e la correzione dei dati in senso longitudinale opera a livello di singolo record attraverso l'identificazione di incoerenze logico-formali che derivano da incompatibilità esistenti in momenti di osservazione diversi. Tali incoerenze possono verificarsi tra valori singolarmente ammissibili a livello trasversale ma anche tra combinazioni di variabili collegate da vincoli logici. Un'ulteriore complicazione nel caso di indagini svolte su individui e famiglie riguarda il fatto che la coerenza sia trasversale che longitudinale deve essere garantita non solo a livello *intra-record* (cioè di singolo individuo) ma anche a livello *inter-record* (cioè tra individui appartenenti alla stessa famiglia).

Quando il processo di correzione è applicato a dati longitudinali, alcuni problemi assumono un aspetto determinante. In primo luogo, i controlli longitudinali e trasversali possono essere effettuati contemporaneamente ma devono essere coordinati e integrati. Generalmente si rende necessario stabilire una priorità tra le regole trasversali e longitudinali che le variabili rilevate devono soddisfare. La coordinazione tra queste due strategie di correzione è importante al fine di evitare il rischio di sovradimensionare il numero di controlli e di modifiche apportate al *data set* originale (Granquist e Kovar, 1997). D'altra parte, nel caso di "nuove" unità rilevate per la prima volta per le quali non sono disponibili dati precedenti, è possibile applicare solo controlli trasversali. Di conseguenza è necessario articolare il processo di controllo e correzione in funzione del numero di interviste precedenti disponibili (che nel caso di EU-SILC, data la struttura del disegno campionario, può variare da 0 a 3).

L'altro aspetto da prendere in considerazione è il fatto che le unità statistiche possono cambiare le loro caratteristiche nel tempo (come lo stato civile, il titolo di studio, l'attività svolta) a seguito di eventi di diversa natura (ad esempio, fusioni o scioglimenti di famiglie, variazioni nei percorsi lavorativi o di studio). Le unità statistiche interessate da questi cambiamenti hanno una loro identità "longitudinale" e l'analisi longitudinale è interessata principalmente alla loro evoluzione nel tempo. Di conseguenza variazioni significative della variabile in oggetto potrebbero non essere il risultato di errori di rilevazione, ma essere causati da cambiamenti effettivi del profilo dell'unità nel tempo. Distinguere tra le due fattispecie non è sempre possibile.

1 I paragrafi 16.1 e 16.6 sono stati redatti da Lucia Coppola; i paragrafi 16.3 e 16.4 sono stati redatti da Daniela Lo Castro; i paragrafi 16.2 e 16.5 sono stati redatti da Mattia Spaziani.

16.2 Strategia di correzione longitudinale in EU-SILC

In un contesto longitudinale, la coerenza di singole osservazioni nel tempo è cruciale perché proprio i percorsi di vita che derivano dall'osservazione ripetuta delle stesse unità sono oggetto dell'analisi longitudinale. I dati di una unità in una data occasione possono essere confrontati con altri valori osservati sulla stessa unità in altri istanti temporali, cioè appartenenti al suo profilo. Bisogna tener conto infatti, non solo dell'esigenza che i percorsi delineati debbano necessariamente seguire delle strutture logiche, evidenziate dai controlli di coerenza e verosimiglianza, ma anche del fatto che alcuni "stati" sono irripetibili.

Alcuni stati, infatti, possono essere occupati una sola volta nella vita: questi cambiamenti di stato sono irreversibili, ovvero il verificarsi in un dato istante della condizione "avere svolto almeno un lavoro" non permette il ritorno alla condizione di "non aver mai lavorato nella vita".

Altre condizioni sono plausibili solo se rispettano determinate sequenze: ad esempio per il titolo di studio è ammesso solo permanere allo stesso livello oppure conseguirne uno superiore; così come non sono ammissibili alcuni cambiamenti di stato civile, come il passare dalla condizione "divorziato" a quella di "separato" con riferimento allo stesso partner.

In altri casi ancora gli stati possono essere occupati più volte nella vita (matrimonio, occupazione): siamo di fronte a singole transizioni dentro e fuori una serie di stati ripetibili.

Le maggiori difficoltà nel trattamento dati longitudinali riguardano le domande relative a informazioni retrospettive (ad esempio l'età di inizio lavoro o di conseguimento del titolo di studio), che sono le più affette da errori legati a problemi di memoria (Kalton *et al.*, 1989). Per questa ragione, dal 2016 una parte di queste incoerenze viene controllata in fase di rilevazione attraverso dei quesiti a conferma su quanto dichiarato nell'occasione di indagine precedente (si veda capitolo 15). Stabilita la validità trasversale delle informazioni raccolte, in presenza di incongruenze longitudinali, le imputazioni/correzioni sono basate sul criterio del minimo cambiamento e/o della prevalenza, implementando degli *edit* longitudinali che sfruttano l'informazione raccolta in tutte le occasioni di rilevazione.

L'individuazione e la correzione degli errori seguendo un approccio deterministico e probabilistico offre la possibilità di realizzare, in tempi diversi, le modifiche necessarie per garantire la coerenza *intra-record*. In tal senso, il trattamento dei dati longitudinali in EU-SILC tradizionalmente ha avuto luogo a valle del trattamento trasversale. Questa strategia era dettata dal fatto che la scadenza per il rilascio dei dati trasversali ad Eurostat fosse antecedente di alcuni mesi rispetto a quella dei dati longitudinali. Dati i tempi di rilascio piuttosto stringenti, si preferiva quindi dare priorità al trattamento trasversale includendo solo alcuni criteri di controllo e correzione basati su informazioni di natura longitudinale; in una fase successiva si aggiungevano ulteriori controlli e correzioni di tipo longitudinale per assicurare una maggior coerenza nei percorsi individuali e familiari osservati e rilasciati nei microdati longitudinali. Di conseguenza, in alcuni casi poteva succedere che il microdato dell'ultimo anno di osservazione longitudinale non coincidesse esattamente con il corrispondente trasversale.

Dall'anno di rilevazione 2014, invece, Eurostat richiede che i dati di entrambe le componenti trasversale e longitudinale vengano forniti congiuntamente per mezzo di un unico file integrato (*reconciled* file). Si è resa quindi necessaria una revisione del processo di controllo e correzione dei microdati al fine di migliorarne al contempo la qualità e la tempestività di rilascio. A fronte di ciò ulteriori procedure di controllo e correzione longitudinale sono state anticipate ed integrate nel processo di creazione dei dati trasversali.

Le principali strategie di trattamento adottate per correggere le incoerenze longitudinali rilevate in EU-SILC possono essere classificate in tre tipologie: (i) correzione "all'indietro", che riproduce il valore dell'intervista più recente sulle precedenti, o vincola l'ammissibilità delle caratteristiche rilevate negli anni precedenti a quelle osservate nell'anno più recente; (ii) correzione "in avanti", che riporta il valore di una intervista passata su quelle successive, o vincola l'ammissibilità delle caratteristiche rilevate nell'anno più recente a quelle già rilevate negli anni precedenti; (iii) correzione secondo la logica della "prevalenza", che imputa il valore più frequente tra quelli rilevati nei diversi anni di osservazione a disposizione.

Nell'indagine EU-SILC è possibile che in una data occasione di indagine l'unità statistica non venga osservata, per cui non sono disponibili le informazioni in tutti e quattro gli anni del panel (perché ad esempio la famiglia è temporaneamente assente e non può essere intervistata). Nell'applicare le diverse strategie di controllo e correzione bisogna quindi tener conto anche dei vuoti informativi dovuti alle mancate interviste: ad esempio, se un individuo è stato intervistato alla prima e alla terza occasione di indagine, ma non alla seconda, si dovranno mettere a confronto le informazioni della terza intervista direttamente con quelle della prima.

16.3 Correzione "all'indietro"

L'implementazione del questionario elettronico rende possibile precaricare una parte dell'informazione rilevata negli anni precedenti, permettendo di definire dei quesiti a conferma e una serie di controlli di coerenza in ottica longitudinale già in fase di raccolta del dato (si veda capitolo 15). Le variabili principalmente interessate sono quelle anagrafiche (data di nascita, sesso, cittadinanza, stato civile, relazione di parentela), quelle relative alle caratteristiche dell'abitazione, al titolo di studio e alla storia lavorativa passata. Poiché l'informazione rilevata viene confermata o rettificata già in fase di intervista, si ritiene più affidabile il dato più recente. Dopo aver verificato la coerenza a livello trasversale nell'ultimo anno di indagine, le correzioni longitudinali seguono per lo più un approccio deterministico, tenendo conto della tipologia di variabile da trattare.

Nel caso più semplice di variabili invarianti nel tempo, che non sono quindi soggette a cambiamenti, come la data di nascita e il sesso, la strategia adottata prevede di replicare il valore dell'ultima intervista disponibile su quelle passate. Più complesso è il caso di quelle variabili che possono seguire solo traiettorie temporali ben definite, come alcune transizioni del livello di istruzione conseguito: in tal caso infatti non ci si può limitare a riportare all'indietro il titolo di studio rilevato nell'ultimo anno, in quanto bisogna preservare i cambiamenti ammissibili in momenti di osservazione contigui, nel rispetto di traiettorie al più crescenti.

A titolo esemplificativo, se nell'anno *t* si è dichiarato di avere il diploma di scuola media superiore ma nell'anno *t-1* una laurea, allora si deve allineare il titolo di studio in *t-1* a quanto osservato in *t*; diversamente, se in *t-1* si era in possesso di una licenza di scuola media inferiore, non è necessario apportare alcuna correzione, poiché il cambiamento osservato per il titolo di studio è ammissibile. Inoltre, anche per titoli di pari livello che possono essere posseduti contemporaneamente senza poter distinguere quale sia quello più alto conseguito, come il master di secondo livello e il diploma di specializzazione universitaria, si preferisce preservare quanto osservato e quindi non si modifica l'informazione raccolta.

In Tavola 16.1 si riporta la distribuzione del titolo di studio per ciascun anno della release longitudinale[2] 2013-2016, prima e dopo la correzione all'indietro delle incoerenze riscontrate per coppie di anni successivi. Il 2016 è l'anno in cui vengono introdotti nel questionario i quesiti a conferma e pertanto è preso come valore di riferimento per il confronto con gli anni precedenti. Osservando le differenze tra le distribuzioni grezza[3] e pulita, si può notare come gli interventi di correzione delle incoerenze tendano a diminuire la frequenza dei titoli più elevati a favore di quelli bassi. L'entità delle incoerenze per ciascuna coppia di anni è mediamente del 7 per cento.

Tavola 16.1 - Distribuzione del titolo di studio prima e dopo le correzioni longitudinali e incidenza delle transizioni incoerenti per coppie di anni. Release longitudinale 2013-2016 (valori per 100 individui di 16 anni e più)

TITOLO DI STUDIO	2013		2014		2015		2016
	grezzo	pulito	grezzo	pulito	grezzo	pulito	grezzo = pulito
Nessun titolo	5,2	6,1	3,4	4,2	2,2	3,0	2,8
Elementare	16,0	16,5	14,3	15,3	14,4	15,2	14,6
Media inferiore	27,3	28,3	27,4	28,6	26,5	27,8	26,9
Media superiore	37,9	36,4	38,9	37,8	39,7	38,4	39,5
Laurea	13,4	12,5	15,6	13,8	16,8	15,3	15,9
Dottorato di ricerca	0,3	0,1	0,4	0,3	0,4	0,2	0,3
Incoerenze tra compresenti in *t-1/t*	**6,4**		**8,4**		**7,1**		**-**

Un altro esempio di correzione all'indietro riguarda lo stato civile. In questo caso le transizioni possibili nello stato civile devono tenere conto, oltre che delle traiettorie plausibili *intra-record*, anche di eventuali cambiamenti nella struttura familiare, in modo da assicurare la coerenza *inter-record* delle tipologie familiari dovute ad eventi quali scioglimenti di unioni o nuovi ingressi in famiglia. È interessante notare che le correzioni che riguardano lo stato civile possono avere un impatto anche sulle relazioni di parentela, che devono essere modificate di conseguenza.

Le incoerenze *intra-record* che prescindono dalle relazioni di parentela con gli altri membri della famiglia sono legate a vincoli più che altro legali. Situazione tipica è il caso in cui l'individuo risulti divorziato nell'anno *t* e celibe/nubile in *t-1*: in questo caso la condizione precedente andrebbe rettificata in separato legalmente, se si tenesse conto delle vecchie norme giuridiche per cui non era possibile ottenere il divorzio in un arco temporale così ristretto; diversamente, è plausibile, e quindi non soggetta a correzione, una transizione del tipo celibe/nubile in *t-1* e separato legalmente in *t* se si ipotizza che l'evento matrimonio (e quindi la condizione di coniugato) sia avvenuto tra le due occasioni di indagine, per quanto non esplicitamente rilevato. A seguito delle recenti disposizioni di legge che normano la disciplina su separazioni e divorzi, sono state aggiornate le possibili transizioni in caso di scioglimento di coppie. Ad oggi è tecnicamente possibile passare da coniugato a divorziato o da celibe a separato di fatto a distanza di un anno di osservazione; per queste situazioni, poiché non è possibile distinguere le transizioni reali dalle errate codifiche, si preferisce non intervenire sui microdati, lasciando così inalterata la transizione.

2 Si veda capitolo 17 per la descrizione della struttura dei file di microdati rilasciati.

3 La variabile grezza qui considerata è quella già corretta a livello trasversale, che quindi soddisfa le regole di compatibilità di ciascun anno; poiché la strategia di correzione longitudinale qui applicata è quella all'indietro, nell'ultimo anno di indagine la variabile grezza coincide con quella pulita.

Per approcciarsi alle correzioni intra-familiari si sono dapprima individuate delle tipologie familiari mettendo a confronto la loro composizione tra un anno di osservazione e il successivo, in modo da individuare per ciascuna di esse la giusta strategia di trattamento: famiglie che non hanno cambiato la loro composizione in due anni di osservazione, famiglie con solo ingressi di nuovi componenti o con solo uscite, famiglie non intervistate in almeno un anno di indagine, altre tipologie (famiglie che hanno sperimentato sia ingressi che uscite, famiglie split). Per valutare l'impatto che l'introduzione delle domande a conferma ha avuto sulle incoerenze nelle transizioni tra stati civili, si riporta in Tavola 16.2 il confronto tra le release longitudinali 2012-2015 e 2013-2016: nel complesso l'incidenza delle incoerenze si è quasi dimezzato passando dal 9 al 4 per cento.

Tavola 16.2 - Distribuzione delle famiglie per variazioni nella composizione familiare e incidenza delle transizioni incoerenti tra stati civili nell'ultima coppia di anni. Release longitudinale 2012-2015 e 2013-2016

TIPOLOGIE FAMILIARI	Release 2012-2015		Release 2013-2016	
	Transizioni 2014/2015		Transizioni 2015/2016	
	Composizione *a livello familiare*	Incoerenze *a livello individuale (compresenti in 2014/2015)*	Composizione *a livello familiare*	Incoerenze *a livello individuale (compresenti in 2015/2016)*
Nessuna variazione	88,3	1,4	85,1	0,4
Solo ingressi	2,9	1,9	2,6	1,2
Solo uscite	5,4	1,2	3,6	0,9
Rientri (a)	2,7	4,5	7,7	1,5
Altre	0,7	-	1,1	-
Totale	**100,0**	**9,0**	**100,0**	**3,9**

(a) Famiglie non intervistate in *t-1* per le quali il confronto è effettuato per la coppia di anni *t-2/t*.

Per le famiglie che hanno mantenuto la stessa composizione, generalmente si impone a tutti i componenti in *t-1* lo stesso valore osservato in *t*, relativamente a relazione di parentela, stato civile, anno di matrimonio e stato civile precedente il matrimonio. L'ipotesi su cui si basa questa strategia è che se la famiglia è composta esattamente dagli stessi individui, non c'è motivo di supporre che le relazioni di parentela e lo stato civile siano cambiati tra un anno e l'altro; poiché nella maggior parte dei casi questi sono stati corretti già in fase di rilevazione tramite le domande a conferma, si considerano più attendibili quelli rilevati nell'anno più recente. Fanno chiaramente eccezione tutti i cambiamenti di stato civile e relazioni di parentela ammissibili, per cui si mantiene l'informazione rilevata nell'anno precedente e quindi il cambiamento osservato tra *t-1* e *t*. Si tratta, ad esempio, degli eventi di matrimonio osservati tra il tempo *t* e *t-1* tra una coppia di conviventi al tempo *t-1*, o dei cambiamenti delle transizioni da separato legalmente a divorziato. In Tavola 16.3 vengono segnalate in rosso le transizioni non ammissibili tra un anno e il successivo, su cui è stata applicata la procedura di correzione all'indietro, che per la release longitudinale 2013-2016 sono pari allo 0,41 per cento.

Nel caso di famiglie dove il nuovo ingresso sia costituito da un coniuge, laddove si osservino transizioni non ammissibili, si copia all'indietro lo stato civile degli individui già presenti in famiglia fatta eccezione per l'individuo campione che si è sposato che mantiene quanto dichiarato l'anno prima. Invece per le famiglie in cui durante il panel si verifica l'uscita di un coniuge dalla famiglia, a seconda che sia avvenuto per decesso o trasferimento, si controlla che lo stato civile del coniuge rimanente sia rispettivamente vedovo o separato/divorziato.

Tavola 16.3 - Transizioni dello stato civile nell'ultima coppia di anni per le famiglie senza variazioni nella composizione familiare. Release longitudinale 2013-2016 *(valori per 100 individui compresenti in 2015/2016)*

2015	2016						
	Celibe o nubile	Coniugato/a	Separato/a di fatto	Separato/a legalmente	Divorziato/a	Vedovo/a	**Totale**
Celibe o nubile	38,29	0,14	0,06	0,03	**0,06**	0,03	**38,61**
Coniugato/a	**0,05**	47,30	**0,02**	**0,02**	**-**	**-**	**47,39**
Separato/a di fatto	**0,03**	**0,01**	1,33	0,04	**-**	0,10	**1,52**
Separato/a legalmente	**0,03**	**0,01**	**0,05**	2,09	0,03	**0,05**	**2,25**
Divorziato/a	**-**	0,02	-	**0,01**	2,27	**0,01**	**2,30**
Vedovo/a	**0,04**	0,02	0,07	**0,01**	**0,01**	7,78	**7,93**
Totale	**38,45**	**47,49**	**1,53**	**2,19**	**2,37**	**7,97**	**100,00**

Nota: Le transizioni incoerenti sono indicate in rosso.

Per le famiglie per cui non è stato possibile effettuare l'intervista nell'anno precedente, si adottano le stesse strategie appena descritte, mettendo a confronto le informazioni delle interviste disponibili più recenti (generalmente quelle osservate due anni prima).

Per le famiglie che hanno sperimentato sia uscite che ingressi di componenti e per le famiglie split, non è possibile individuare una procedura generalizzata che ne permetta una specifica correzione, in quanto la complessità dei controlli intra-familiari da effettuare è tale da costringere ad analizzare manualmente caso per caso e decidere di volta in volta la correzione più adatta. Si tratta tuttavia di pochi casi, circa l'1 per cento (Tavola 16.2).

16.4 Correzione "in avanti"

Nella seconda casistica di correzioni longitudinali rientrano le variabili per le quali il passaggio di stato determina una permanenza definitiva in quello stesso stato, una volta raggiunto il quale non è più possibile tornare alla condizione precedente. Si tratta per lo più di variabili dicotomiche, che indicano una situazione di presenza/assenza di un particolare attributo. In questi casi l'anno su cui intervenire per apportare la correzione è il più recente tra quelli osservati e viene imputato in relazione al valore assunto nelle precedenti occasioni.

Un esempio tipico è l'avere mai lavorato nella vita: la prima volta che si dichiara che c'è stato almeno un episodio lavorativo nella propria vita, allora si deve permanere sempre in tale condizione. L'assegnazione del valore da riportare in avanti sfrutta tutte le informazioni trasversali e longitudinali a disposizione: si individuano infatti segnali di occupazione passata e attuale e di reddito da lavoro percepito, che aiutano a determinare se l'individuo abbia mai lavorato nella vita, e in funzione di ciò si corregge l'eventuale incoerenza.

Nello specifico si identificano per ogni coppia di anni i percorsi ammissibili dello stato occupazionale tra *t-1* e *t* e, a seconda della presenza o meno di segnali oggettivi di reddito da lavoro in almeno uno degli anni del panel, si effettuano le correzioni in avanti nel primo caso o all'indietro nel secondo. In Tavola 16.4 si riporta l'impatto delle correzioni longitudinali per i casi di incoerenza riscontrati in ogni coppia di anni della release longitudinale 2013-2016. L'entità delle correzioni in tutte le coppie di anni è contenuto: nel complesso gli interventi hanno interessato circa il 2 per cento dei record (righe in rosso), coinvolgendo in misura maggiore le correzioni all'indietro solo per la prima coppia di anni.

Tavola 16.4 - Transizioni dello stato occupazionale tra coppie di anni consecutivi per tipo di percorso, presenza/assenza di segnali di occupazione e di reddito e tipo di correzione. Release longitudinale 2013-2016 (valori per 100 individui di 16 anni o più compresenti in *t-1/t*)

Tipo di percorso	Segnale di occupazione		Segnale di reddito in *t* o *t-1*	Tipo di correzione	Transizioni *t-1/t*		
	t-1	*t*			2013/2014	2014/2015	2015/2016
coerente	sì	sì	-	-	82,9	83,1	84,7
coerente	no	no	-	-	12,3	12,4	11,4
coerente	no	sì	-	-	2,8	2,5	2,1
incoerente	**sì**	**no**	**no**	**all'indietro**	**1,3**	**1,0**	**0,6**
incoerente	**sì**	**no**	**sì**	**in avanti**	**0,8**	**1,0**	**1,3**

Una volta che si modifica in *t* la condizione da "non ha mai lavorato" a "ha lavorato almeno una volta" nella vita, le altre variabili della sezione lavoro che da essa dipendono vengono donate sulla base di quanto dichiarato in *t-1*.

Anche per questa variabile è previsto l'ausilio della domanda a conferma in fase di intervista; tuttavia in generale il controllo temporale è effettuato rispetto al solo anno precedente se presente, o al più su due anni prima. Questo non mette al riparo da eventuali incoerenze che si possono comunque verificare nel momento in cui si osserva la traiettoria completa nei quattro anni: infatti in questo caso specifico la natura dicotomica della variabile impone che, al manifestarsi per la prima volta della condizione, questa venga mantenuta per tutte le occasioni successive (per cui se alla quarta intervista viene confermato il valore rilevato nella terza, ma esso risulta incompatibile con quanto rilevato durante la prima intervista, allora si userà comunque il valore della prima intervista per correggere le successive).

16.5 Correzione basata sulla "prevalenza"

L'approccio basato sulla logica della prevalenza mira a imputare i casi di incoerenza o di valori mancanti assegnando la modalità che si presenta con maggior frequenza nel periodo di osservazione. Questo intervallo temporale ha una durata variabile tra due e quattro anni di indagine ed è condizionato a eventuali vincoli di coerenza trasversale. In questa tipologia di correzione rientrano il trattamento di variabili quantitative come l'età in cui si è iniziato a lavorare e l'anno di conseguimento del titolo di studio più elevato.

Attenzione particolare merita quest'ultimo esempio, in quanto in tal caso bisogna tener conto simultaneamente di controlli di coerenza trasversale legati sia all'età anagrafica che al livello di istruzione compatibile con quell'età. In linea di principio, più è ampio l'intervallo temporale tra la data in cui si è verificato l'evento e quella in cui si chiede di ricordarlo, tanto maggiore è lo sforzo di memoria richiesto. In particolare, essendo un evento occasionale a volte molto lontano nel tempo e del quale si richiede l'anno esatto, i casi di incoerenze sono molto frequenti.

Nell'effettuare le correzioni longitudinali, bisogna considerare innanzitutto che il valore dell'anno di conseguimento del titolo va mantenuto costante se il titolo di studio rimane invariato, oppure va equiparato all'età anagrafica laddove si osserva nel corso dell'indagine il conseguimento di un titolo di studio superiore. Una volta individuati i casi di incoerenza, si calcola il valore modale sui dati osservati, da assegnare agli anni da correggere; nel caso di soli due anni di osservazione o di coppie di anni uguali, per cui non è possibile individuare quale dei due valori è quello corretto, si è deciso di considerare come più attendibile quello più recente. Questa strategia ha assunto maggiore fondatezza con l'introduzione della domanda a conferma

sul titolo di studio, che rende più affidabile l'informazione più recente, ovvero quella confermata o rettificata in fase di rilevazione. Infine, nelle situazioni di incertezza maggiori dove non è possibile determinare il valore più frequente, si individua come candidato da donare per l'imputazione il valore tra quelli effettivamente osservati che ha distanza minima dal valore mediano.

Questa procedura assicura la coerenza longitudinale dell'anno di conseguimento del titolo di studio in termini di progressione temporale, nel tentativo di salvaguardare il più possibile l'informazione rilevata in sede di intervista. Selezionando il valore più frequente si cerca quindi di limitare eventuali errori di memoria attraverso l'individuazione della serie corretta a partire da almeno uno dei valori realmente osservati. Si fa notare che l'introduzione delle domande a conferma è in grado di ridurre l'impatto del numero di correzioni effettuate con l'ottica della prevalenza; infatti, se confermato, il valore diventa implicitamente quello prevalente in tutti i panel, poiché almeno i valori degli ultimi due anni coincidono. Si rende comunque necessario applicare questa strategia per tutte quelle casistiche in cui il dato rilevato viene modificato dalle procedure di controllo e correzione di tipo trasversale, per cui la prevalenza "implicita" viene a mancare.

In Tavola 16.5 si riportano le distribuzioni dell'anno di conseguimento del titolo di studio prima e dopo le correzioni longitudinali per i compresenti nei quattro anni del panel 2013-2016. Le correzioni hanno riguardato principalmente l'ultimo anno del panel, sia in termini di indici di posizione che di variabilità, a differenza dei primi tre anni in cui non si evidenziano variazioni significative. Questo dipende da due fattori.

In primo luogo dal modo di costruzione del file di microdati longitudinale, realizzato unendo le quattro base dati trasversali. La scelta di quali file iniziali utilizzare è ricaduta su quella dei dati che hanno già subito un primo piano di controlli e correzioni trasversali, anziché ripartire di volta in volta da quelli rilevati, ovvero non ancora trattati.

L'altro aspetto da tenere in considerazione è l'avere invertito il tipo di correzione del titolo di studio, con cui deve essere rispettata la coerenza dell'anno di conseguimento del titolo stesso: fino alla release longitudinale 2012-2015 il titolo veniva corretto in avanti, dalla release 2013-2016 l'introduzione dei quesiti a conferma ha determinato la scelta di effettuare le correzioni all'indietro.

La combinazione di questi due fattori fa sì che le distribuzioni osservate nei primi tre anni risentano delle correzioni già apportate sui microdati trasversali, "trascinando" sull'ultimo anno l'impatto maggiore delle correzioni.

Tavola 16.5 - Distribuzione dell'anno di conseguimento del titolo di studio prima e dopo le correzioni longitudinali. Compresenti nei 4 anni del panel 2013-2016

ANNO DI CONSEGUIMENTO DEL TITOLO	10° Percentile	Primo Quartile	Mediana	Terzo Quartile	90° Percentile	Differenza Interquartile
Anno 2013						
Grezzo	1955	1967	1982	1997	2008	30
Pulito	1955	1967	1982	1997	2008	30
Anno 2014						
Grezzo	1955	1968	1984	2001	2012	33
Pulito	1955	1968	1985	2009	2014	41
Anno 2015						
Grezzo	1956	1970	1989	2014	2014	44
Pulito	1956	1969	1987	2012	2014	43
Anno 2016						
Grezzo	1955	1967	1982	1997	2009	30
Pulito	1956	1970	1987	2014	2014	44

16.6 Conclusioni

Le strategie di correzione dei dati individuate per il trattamento della componente longitudinale dell'indagine EU-SILC rispettano il tipo di informazioni a disposizione in fase di rilevazione, la tipologia delle traiettorie degli stati, la natura della variabile.

La presenza dei quesiti a conferma induce a considerare come più attendibile l'informazione rilevata nell'ultimo anno: per questo l'approccio utilizzato è quello di apportare le correzioni di eventuali incoerenze "all'indietro", riportando il valore dell'ultimo anno sui precedenti.

Bisogna tenere conto anche dell'esistenza di percorsi ammissibili che in alcuni casi possono rispettare solo determinate sequenze: nel caso in cui il passaggio da una condizione all'altra non permette più il ritorno alla condizione precedente, la strategia adottata è quella di correggere "in avanti" le transizioni incoerenti.

Nel caso di variabili quantitative, la disponibilità di osservazioni ripetute nel tempo sullo stesso individuo permette di determinare il valore più coerente come quello più frequente: in questo caso si implementa la logica della "prevalenza".

Il file dei microdati longitudinale è costruito unendo i quattro anni del panel, partendo dai dati già soggetti a un primo piano di controlli e correzioni trasversali per quanto riguarda l'ultimo anno del panel e trattati anche longitudinalmente nel caso dei primi tre anni. Questo può rappresentare un limite per le correzioni longitudinali "in avanti" o basate sulla "prevalenza" perché appiattisce le transizioni "trascinando" sull'ultimo anno le informazioni delle interviste precedenti, senza sfruttare il dato longitudinale originario effettivamente rilevato. D'altra parte il dato rilevato è penalizzato dal non essere stato trattato in alcun modo e quindi ancora soggetto alle incoerenze trasversali: ciò richiederebbe l'implementazione di un piano di *check* che integri regole di compatibilità sia trasversali che longitudinali, da applicare simultaneamente ai dati appartenenti a quattro anni di indagine; è facilmente intuibile come lo sviluppo di una tale procedura sia di difficile realizzazione.

17. L'UTILIZZO DEI DATI IN OTTICA LONGITUDINALE: POTENZIALITÀ DEL DISEGNO CAMPIONARIO A GRUPPI ROTAZIONALI[1]

17.1 Introduzione

L'indagine EU-SILC, in Italia così come nella maggior parte dei Paesi che partecipano alla rilevazione, si basa su un campione ruotato composto da 4 gruppi rotazionali o panel, ciascuno dei quali viene osservato per quattro anni consecutivi. Questa articolata struttura campionaria consente un uso flessibile dei dati per produrre stime sia trasversali sia longitudinali. Inoltre, nell'ambito delle stime longitudinali, ci si può avvalere delle unità campionarie appartenenti a uno o più panel, a seconda degli obbiettivi conoscitivi.

In questo capitolo si vuole fornire una lettura integrata e sintetica delle caratteristiche dei campioni longitudinali e delle popolazioni di cui sono rappresentativi, e offrire delle linee guida per la scelta dei panel, delle unità campionarie e dei pesi di riporto all'universo più adeguati secondo le esigenze conoscitive. In particolare, nel paragrafo 17.2 si descrivono i campioni e le popolazioni longitudinali, nel paragrafo 17.3 il sistema di pesi longitudinali stimati e il loro utilizzo, e nel paragrafo 17.4 si discutono le principali conclusioni.

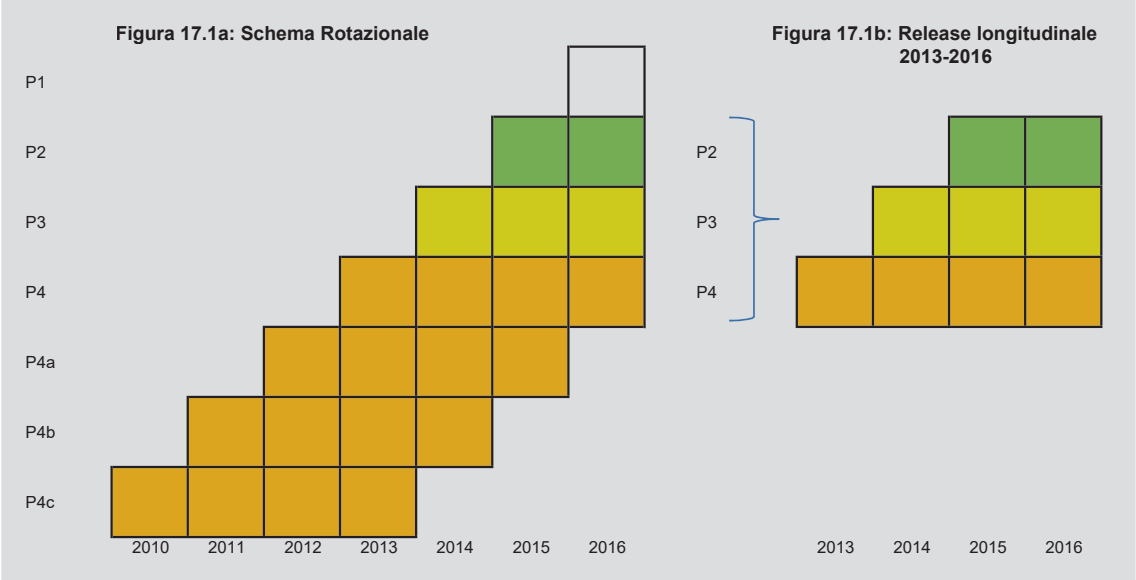
17.2 Campioni e popolazioni longitudinali

L'implementazione dello schema rotazionale di EU-SILC viene mostrato nella Figura 17.1a, che a titolo esemplificativo fa riferimento solo agli anni dal 2010 al 2016, benché al momento sia disponibile una serie storica più ampia di panel completi (cioè osservati per 4 anni consecutivi), a partire da quella del 2004-2007. Ogni anno l'Istat rilascia una release longitudinale, composta dai tre panel più recenti (osservati per due, tre o quattro anni consecutivi) come indicato in Figura 17.1b.

Ciascun panel è indipendente ed è rappresentativo della popolazione "trasversale" dell'anno di estrazione del campione e della popolazione "sopravvivente" negli anni successivi. Ovvero, come mostrato nella Figura 17.2, con riferimento ad esempio al panel intervistato per 4 anni (P4), gli individui che fanno parte di famiglie intervistate il primo anno sono rappresentativi della popolazione trasversale 2013 sia in termini di individui sia di famiglie (60.486.301 individui e 25.554.875 famiglie). Così come per il campione trasversale, la popolazione del 2013 è stimata come la popolazione residente in famiglia al 31 dicembre 2012, cui si aggiungono i nuovi nati tra il primo gennaio 2013 e il momento dell'intervista, secondo quanto risulta dalla rilevazione.

1 I paragrafi 17.1 e 17.2 sono stati redatti da Lucia Coppola; il paragrafo 17.4 è stato redatto da Daniela Lo Castro; il paragrafo 17.3 è stato redatto da Mattia Spaziani.

Figura 17.1 - Schema rotazionale e release longitudinale



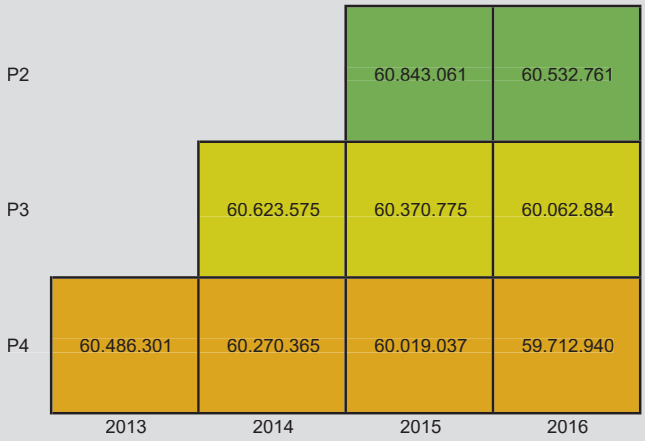
In occasione della prima intervista vengono definiti gli individui campione, ovvero gli individui di almeno 14 anni che dovranno continuare a far parte del panel nei tre anni successivi, anche in caso di trasferimento nell'ambito del territorio nazionale. Si definiscono anche le famiglie campione, ovvero le famiglie in cui sia presente almeno un individuo campione. In queste famiglie vengono rilevati non solo gli individui campione, ma tutti i componenti di fatto presenti nei diversi anni di rilevazione. Gli individui che alla prima intervista hanno meno di 14 anni e gli individui che entrano a far parte delle famiglie campione a partire dalla seconda intervista in poi, vengono definiti coresidenti. Se i coresidenti escono da una famiglia campione, ovvero si trasferiscono nell'ambito del territorio nazionale ma senza altri individui campione, non vengono più ricontattati ed escono dal panel. Inoltre, gli individui che entrano nella rilevazione dopo la prima intervista, se sono eleggibili (devono avere almeno 16 anni al 31 dicembre precedente l'intervista) vengono intervistati e contribuiscono alle stime di tipo trasversale, ma non a quelle di tipo longitudinale, in quanto non facendo parte del campione di partenza, non sono necessariamente rappresentativi della popolazione estratta di cui il panel è invece rappresentativo. Per questo motivo, questi individui e le loro caratteristiche sono presenti nella release longitudinale ma viene assegnato loro un peso longitudinale nullo[2].

Per definire il campione da rilevare nella seconda intervista (e nelle successive), si applicano le regole di inseguimento definite dal Regolamento Europeo (European Commision, 2003). Il campione teorico della seconda intervista è composto dagli individui di famiglie intervistate nell'anno precedente, al netto di coloro che escono dalla popolazione *target* [3], per decesso, trasferimento in convivenza o all'estero. L'ammontare della popolazione longitudinale del 2014 (panel P4 in Figura 17.2) è stimata a partire dalla popolazione del 31 dicembre 2012, cui si sottrae il saldo naturale e i cancellati per l'estero osservati nel 2013 (secondo le stime ufficiali dell'Istat, http://demo.istat.it/) per ottenere la popolazione

2 Questa strategia riguarda solo i pesi longitudinali. Nella release trasversale, infatti, anche gli individui che entrano a far parte delle famiglie campione dopo la prima intervista, ricevono il peso della famiglia.

3 La popolazione oggetto di indagine è quella residente nel territorio nazionale e che vive in famiglia.

Figura 17.2 - Schema rotazionale e popolazioni longitudinali: somma dei pesi individuali (RB060) per panel ed anno



longitudinale al 31 dicembre 2013, e si aggiungono i nuovi nati tra il 1 gennaio 2014 e il momento dell'intervista 2014, secondo quanto stimato dalla rilevazione. La stima dei nuovi nati si ottiene attribuendo loro il peso della madre (come indicato nelle linee guida fornite da Eurostat) (European Commission, 2017). Ne consegue che il panel che nel 2014 è alla seconda intervista rappresenta un totale di popolazione diverso da quello che nel 2014 è alla prima intervista. Nel primo caso si tratta della popolazione trasversale del 2013 "sopravvivente" al 2014, mentre nel secondo caso si tratta della popolazione trasversale del 2014.

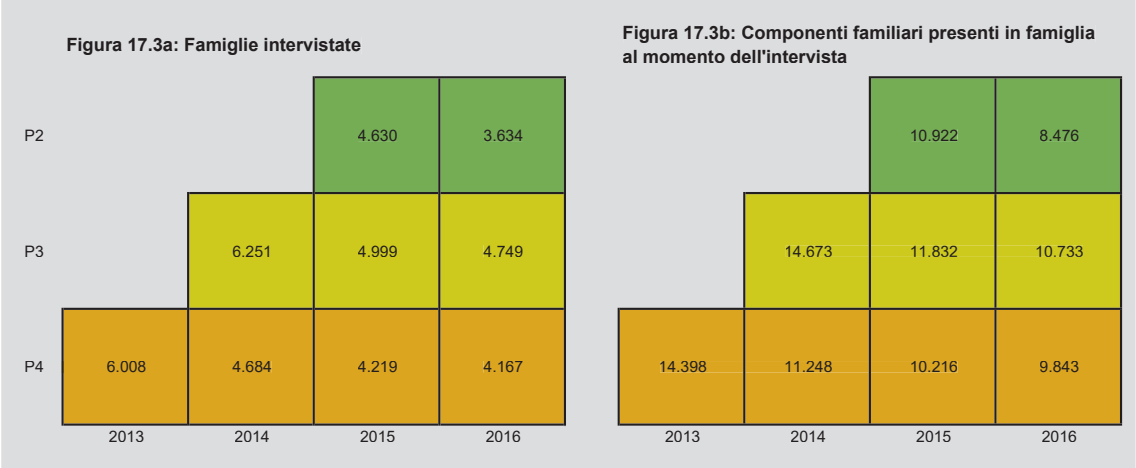
Per definire le popolazioni della terza e quarta intervista, si segue la stessa strategia utilizzata per la seconda, e l'ammontare della popolazione longitudinale stimato per ciascun panel e ciascun anno è mostrato in Figura 17.2.

Oltre ai movimenti naturali della popolazione (nascite, decessi e trasferimenti) di cui il panel vuole essere rappresentativo, bisogna tener conto della possibile perdita di unità campionarie per motivi che esulano dai cambiamenti osservati sulla popolazione (*attrition*) e dovuti ad esempio a: difficoltà nel ricontattare le famiglie durante il periodo di rilevazione, soprattutto nei casi di trasferimento sul territorio nazionale; mancanza di disponibilità della famiglia a partecipare alla rilevazione per impedimenti temporanei, dovuti ad esempio a condizioni di salute; rifiuto definitivo da parte della famiglia di partecipare alla rilevazione. Nelle interviste successive alla prima, in caso di mancato contatto con la famiglia per motivi temporanei, in conformità con il Regolamento Europeo (European Commision, 2003), si procede con un nuovo tentativo di contatto anche nell'anno successivo. Nel caso di rifiuto a collaborare, invece, la famiglia non viene più ricontattata.

La combinazione dell'uscita delle unità campionarie dai panel, sia per motivi naturali, sia per effetto dell'*attrition*, fa sì che le dimensioni campionarie siano variabili a seconda dell'anno e del panel. Nella Figura 17.3, ad esempio, vengono mostrate le numerosità delle famiglie intervistate e dei relativi componenti familiari, per anno e panel. Si nota che per il panel completo (P4) le famiglie intervistate, da 6.008 alla prima intervista, si sono ridotte a 4.167 della quarta (analogamente i componenti familiari passano da 14.398 a 9.843). È importante notare che la dimensione campionaria alla prima intervista è anch'essa soggetta a variabilità, poiché dipende dai tassi di mancata risposta realizzati in quello specifico anno di rilevazione. Il panel estratto nel 2015, ad esempio, ha una dimensione campionaria iniziale decisamente ridotta rispetto ai precedenti panel (4.630 famiglie e 10.922 compo-

nenti familiari), a causa di un maggior tasso di caduta realizzatosi durante il periodo di rilevazione del 2015.

Figura 17.3 - Numerosità delle famiglie intervistate e dei componenti familiari, per panel e anno



La complessa articolazione delle regole di inseguimento dà luogo a molteplici percorsi di presenza/assenza degli individui durante il periodo di osservazione del panel. Nella Figura 17.4 sono rappresentati tutti i possibili percorsi osservabili nei panel, a seconda del numero di interviste effettuate (si indica la presenza dell'individuo con P e la sua assenza con --). Quindi, oltre agli individui che vengono rilevati in tutti gli anni di interesse (ad esempio riga 7 nel panel a 4 anni), ci sono quelli che durante il panel escono definitivamente dall'osservazione (ad esempio righe 11, 13 e 14 per il panel a 4 anni), e quelli che mostrano una presenza intermittente (ad esempio 8, 9, 10 e 12 nel panel a 4 anni).

Figura 17.4 - Percorsi di presenza (P) e assenza (--) degli individui, per panel ed anno e numerosità campionaria

	2013	2014	2015	2016	Riga	Numerosità campionaria
P2			P	P	1	8367
			P	--	2	2555
P3		P	P	P	3	9205
		P	--	P	4	1215
		P	P	--	5	2437
		P	--	--	6	1816
P4	P	P	P	P	7	8071
	P	P	--	P	8	854
	P	--	P	P	9	450
	P	--	--	P	10	2
	P	P	P	--	11	1293
	P	--	P	--	12	68
	P	P	--	--	13	849
	P	--	--	--	14	2811

Nota: il peso longitudinale RB060 è definito per tutti gli individui e per tutti gli anni contrassegnati dalla lettera P; il peso longitudinale RB064 è definito per l'anno 2016 per tutti gli individui della riga 7; il peso longitudinale RB063 è definito per l'anno 2016 per tutti gli individui delle righe 7 e 3; il peso longitudinale RB062 è definito per l'anno 2016 per tutti gli individui delle righe 9, 7, 3 e 1.

La perdita di unità campionarie può avere degli effetti sulle stime non solo perché la dimensione campionaria diminuisce, ma soprattutto perché le unità che rimangono nel cam-

pione possono avere caratteristiche diverse rispetto a quelle che ne escono e quindi essere rappresentative solo di un sottoinsieme selezionato della popolazione (*attrition* selettivo). Per ridurre gli effetti di questa possibile distorsione, l'Istat produce un articolato sistema di pesi che tiene conto della selezione della mancata risposta tra il primo anno di intervista e i successivi, in modo tale che le unità campionarie che restano nel campione siano il più possibile rappresentative della popolazione di partenza.

17.3 Sistema di pesi longitudinali e strategia di stima

Secondo quanto previsto dal Regolamento Europeo, la release longitudinale di EU-SILC include un sistema di pesi che possono essere usati in alternativa a seconda degli obbiettivi di stima. In particolare, per ciascun panel e ciascun anno vengono rilasciati dei pesi che rappresentano le caratteristiche della popolazione longitudinale, ovvero la popolazione della prima occasione di indagine e sopravvivente negli anni successivi, il cui ammontare, per la release longitudinale 2013-2016, è rappresentato nella Figura 17.2. Tali pesi sono denominati RB060 e sono forniti per tutti gli individui presenti in ciascuna occasione di indagine (indicati con P nella Figura 17.4). Tramite l'uso di questi pesi, quindi, si può descrivere la popolazione longitudinale e la sua evoluzione del periodo di osservazione, in termini ad esempio di struttura per sesso ed età, o di individui a rischio di povertà ecc.

Inoltre, vengono rilasciati, solo per l'ultimo anno della release longitudinale, dei pesi relativi agli individui "compresenti" negli ultimi due, tre o quattro anni di osservazione (ovvero gli individui che risultano presenti in famiglia intervistate in tutti gli anni presi in considerazione), denominati rispettivamente RB062, RB063 e RB064, che possono essere utilizzati per la stima delle transizioni che hanno avuto luogo nel periodo di interesse.

17.3.1 Pesi longitudinali per panel e anno di osservazione (RB060)

I pesi longitudinali alla prima intervista coincidono con quelli trasversali, riproporzionati in modo tale che il sottocampione del panel alla prima intervista sia rappresentativo dell'intero campione trasversale del medesimo anno di rilevazione. Quindi alla prima intervista tutti i componenti familiari hanno lo stesso peso che coincide con quello familiare (come brevemente descritto nel capitolo 12, paragrafo 12.5).

Durante il periodo di osservazione del panel le famiglie possono cambiare struttura e dimensione. Inoltre, gli individui campione possono lasciare le famiglie di origine per formarne di nuove (denominate famiglie split). È opportuno quindi considerare l'individuo come unità di rilevazione longitudinale, da inseguire nel tempo e sul territorio nazionale. Ne consegue che anche la strategia di stima dei pesi, che include la correzione per mancata risposta tra una intervista e l'altra, sia sviluppata a livello individuale. Quindi, dalla seconda intervista in poi, ciascun componente familiare ha un proprio peso longitudinale, che può essere diverso da quello degli altri componenti familiari. Per compensare i possibili effetti selettivi dell'*attrition*, vengono stimati dei modelli di mancata risposta a livello individuale tra la prima intervista e ciascuna delle successive, che tengono conto di diverse caratteristiche sia individuali sia familiari.

Formalmente, il peso dell'individuo *j* all'intervista *t* è dato da $\hat{p}_{jt} = p_{j1}\frac{1}{\gamma_{jt}}(1-\beta_{jt})$, dove p_{j1} è il peso del medesimo individuo alla prima intervista, γ_{jt}

è la probabilità di risposta all'intervista t e β_{jt} è la probabilità di essere uscito dalla popolazione *target* (cioè essere deceduto, trasferito all'estero o in convivenza) all'intervista t.

La stima della probabilità di risposta γ_{jt} e di uscita dalla popolazione *target* β_{jt} sono ottenute tramite modelli logistici che tengono conto delle seguenti caratteristiche rilevate alla prima intervista: sesso, classi di età, ripartizione geografica, dimensione demografica del comune, livello di istruzione, condizione professionale auto-dichiarata, fonte principale di reddito, titolo di godimento dell'abitazione, numero di sintomi di deprivazione, quinto di reddito equivalente, numero di componenti della famiglia.

Agli individui che entrano a far parte del panel a partire dalla seconda intervista viene attribuito un peso nullo, con l'eccezione dei nuovi nati cui, invece, viene attribuito il peso della madre.

Per ciascun panel e ciascun anno, quindi, viene corretto il peso assegnato alla prima intervista, in modo tale da ridurre gli effetti selettivi dell'*attrition*, per lo meno secondo le caratteristiche individuali e familiari prese in considerazione. Infine, si applica un passo di calibrazione al totale della popolazione longitudinale (stimato così come indicato nel paragrafo 17.2) e alle transizioni tra condizioni occupazionali[4] in coppie di anni stimate dalla rilevazione sulle Forze di lavoro[5].

Ad esempio, per la stima dei pesi longitudinali da attribuire agli individui appartenenti al panel di quattro anni (P4 in Figura 17.4) e presenti nel 2016, la probabilità di mancata risposta viene stimata mettendo a confronto gli individui intervistati nel 2013 ma non nel 2016 (righe 11, 12, 13 e 14) con tutti gli individui intervistati sia nel 2013 che nel 2016 (righe 7, 8, 9 e 10), tenendo conto delle loro caratteristiche individuali e familiari alla prima rilevazione. Il peso attribuito a ciascun individuo alla prima intervista (in questo caso nel 2013) viene quindi aggiustato secondo una specifica probabilità di mancata risposta, associata alle caratteristiche iniziali dell'individuo. Pertanto, gli individui che appartengono alla stessa famiglia, pur avendo lo stesso peso alla prima intervista, avranno pesi diversi negli anni successivi. In questo modo, le unità di rilevazione che rimangono nel panel diventano rappresentative anche di quelle che ne sono uscite e che avevano le stesse caratteristiche iniziali. Ovviamente la correzione per mancata risposta può tenere conto di un numero limitato di caratteristiche, per cui sono state scelte quelle più rilevanti per gli obbiettivi di indagine. Se i dati longitudinali di EU-SILC vengono usati per lo studio di fenomeni particolari, associati a caratteristiche non prese in considerazione, i modelli di correzione dei pesi implementati potrebbero non essere sufficienti e l'utente potrebbe aver bisogno di utilizzare ulteriori strategie per controllare possibili effetti di selezione dovuti all'*attrition*.

17.3.2 Pesi per la stima di indicatori su unità compresenti (RB062, RB063 e RB064)

Qualora sia necessario condurre le analisi su panel bilanciati, ovvero solo sul sottoinsieme di unità compresenti in più anni consecutivi, è utile adattare i pesi fino ad ora discussi in modo tale da rendere le sole unità di analisi compresenti rappresentative delle popolazioni

4 Si fa riferimento alla classificazione ILO che viene rilevata ma non viene rilasciata nella release longitudinale, che, come da Regolamento Europeo, include invece la condizione occupazionale auto-dichiarata.

5 Si includono anche vincoli su alcune caratteristiche della popolazione longitudinale stimate a seguito della correzione dei pesi per mancata risposta, in modo tale che non vengano alterate dall'introduzione dei vincoli sul totale di popolazione e sulle transizioni occupazionali: struttura della popolazione per sesso e classi di età (0-15; 16-24; 25-44; 45-64; 65+), ripartizione geografica di residenza (NUTS 1), e indicatori di maggiore interesse (rischio di povertà, deprivazione materiale (grave e non), bassa intensità lavorativa e rischio di povertà o esclusione sociale).

longitudinali di interesse. Inoltre, è possibile sfruttare l'articolazione della struttura campionaria di EU-SILC in modo flessibile, unendo le osservazioni campionarie appartenenti a diversi panel, laddove sia utile.

Ad esempio, con riferimento ai dati della release longitudinale 2013-2016 fino ad ora discussi, vengono prodotti e rilasciati: (i) i pesi che consentono di ottenere stime sulla popolazione estratta nel 2013 e sopravvissuta fino al 2016, basandosi sui soli compresenti nei 4 anni (denominati RB064)[6]; (ii) i pesi che consentono di ottenere stime sulla popolazione estratta nel 2014 e sopravvissuta fino al 2016, basandosi su tutte le unità campionarie compresenti nei 3 anni, appartenenti ai due panel intervistati in questi anni (denominati RB063)[7]; (iii) i pesi che consentono di ottenere stime sulla popolazione estratta nel 2015 e sopravvissuta nel 2016, basandosi su tutte le unità campionarie compresenti nei 2 anni, appartenenti a tutti e tre i panel intervistati in questi anni (denominati RB062)[8].

Di seguito vengono presentati degli esempi per chiarire quando e perché è opportuno usare questi pesi.

Persistenza in povertà

L'indicatore longitudinale di rilevanza centrale, tra quelli definiti da Eurostat, è il rischio di povertà persistente (Persistent at Risk of Poverty Rate), secondo il quale si definiscono persistentemente poveri gli individui che sono a rischio di povertà nell'ultimo anno di rilevazione e almeno in due dei tre anni precedenti (Eurostat, 2014, pp: 269-273).

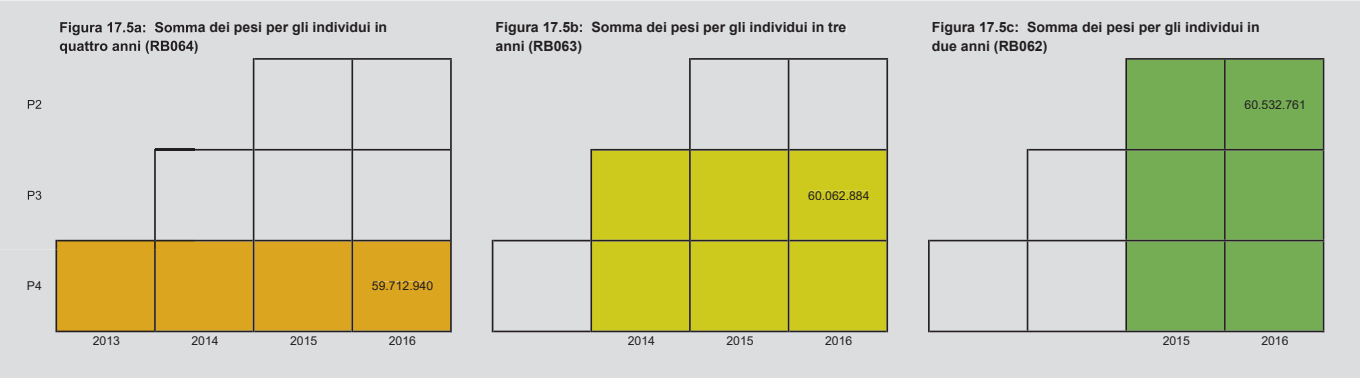
Per avere le informazioni necessarie per la stima di questo tipo di indicatori è necessario conoscere le caratteristiche delle unità campionarie in tutti e quattro gli anni di osservazione. Le stime, quindi, si basano solo sulle unità compresenti nei quattro anni (riga 7 in Figura 17.4, corrispondente nell'esempio a 8.071 unità). È necessario, quindi, avere un peso che consenta di rappresentare la popolazione longitudinale usando solo questo sottoinsieme di unità (denominato RB064 nella release longitudinale). Per ottenere questo peso, si parte da quello già stimato per tutte le unità del panel giunto alla quarta intervista presenti nell'ultimo anno di rilevazione (cioè RB060 del 2016 per gli individui nelle righe 7, 8, 9 e 10 nella Figura 17.4), si impongono uguali a 0 i pesi degli individui che per almeno un anno non hanno partecipato alla rilevazione (righe 8, 9 e 10 di Figura 17.4), si riproporzionano i pesi degli individui compresenti (riga 7 in Figura 17.4) in modo che siano rappresentativi dell'intera popolazione longitudinale (per cui la somma dei pesi RB064 coincide con la somma dei pesi RB060 nel 2016 del panel giunto alla quarta intervista, come in Figura 17.5a). Questa strategia consente di ottenere una stima, sia percentuale che assoluta, dei persistentemente poveri, pur basandosi solo sulle unità campionarie compresenti.

Ricorrere a un riproporzionamento per ottenere i pesi per le sole unità compresenti equivale a ipotizzare che la probabilità di mancata risposta degli individui non presenti nel 2016 (ovvero quella stimata per ottenere i pesi RB060 del 2016) e di quelli presenti nel 2016 ma non presenti in uno degli anni precedenti (righe 8, 9 e 10 in Figura 17.4) sia analoga. Se le due tipologie di non rispondenti fossero significativamente diverse sarebbe necessario stimare un modello di mancata risposta ad-hoc, per tutti coloro che almeno una volta durante il panel non sono stati intervistati. Entrambe le possibilità sono state testate

6 Individui rappresentati in riga 7 in Figura 17.4.
7 Individui rappresentati in riga 7 e 3 in Figura 17.4.
8 Individui rappresentati in riga 9, 7, 3 e 1 in Figura 17.4.

sui dati italiani di EU-SILC, e poiché il ricorso al modello di mancata risposta ad-hoc fornisce stime analoghe a quelle che si ottengono con il semplice riproporzionamento, per semplicità si è preferito seguire questa seconda strategia.

Figura 17.5 - Individui compresenti in 4, 3, e 2 anni e somma dei pesi longitudinali (RB064, RB063 e RB062)



Gli indicatori che sintetizzano quattro anni di osservazione, come il numero di anni trascorsi in povertà o il rischio di povertà persistente, presentano il vantaggio di sfruttare completamente il periodo di osservazione dell'indagine, sintetizzando dinamiche di più lungo periodo. D'altra parte, presentano lo svantaggio di utilizzare un numero relativamente esiguo di unità campionarie (poco più di 8 mila nell'esempio), che può variare di anno in anno, poiché i livelli di *attrition* nelle varie occasioni d'indagine sono anch'essi variabili.

Transizioni tra coppie di anni

Data la struttura campionaria di EU-SILC, se gli obbiettivi di analisi sono circoscritti ad un periodo di osservazione più limitato, come ad esempio le transizioni osservate in coppie di anni, si può ricorrere a tutte le unità di rilevazione osservate nell'ultima coppia di anni, anche se appartenenti a panel diversi. In altre parole, se si vuole stimare una transizione tra il 2015 e il 2016, anziché selezionare i compresenti appartenenti solo al panel alla seconda intervista (riga 1 in Figura 17.4), si possono includere nell'analisi anche le altre unità compresenti nei due anni, appartenenti agli altri due panel (riga 3 per il panel alla terza intervista e righe 7 e 9 per il panel alla quarta intervista, in Figura 17.4). Questa strategia consente di basare le stime su 26.093 unità anziché le sole 8.367 del panel alla seconda intervista.

Poiché, come si è detto, ciascun panel è rappresentativo di una diversa popolazione longitudinale, ovvero quella dell'anno di estrazione e sopravvissuta negli anni successivi, si dovranno adattare i pesi già prodotti indipendentemente per ciascun panel (RB060), in modo che siano rappresentativi della stessa popolazione, nel caso specifico di quella estratta nel 2015 e sopravvissuta nel 2016. Ovvero, si impone che tutte le unità campionarie osservate tra il 2015 e il 2016 (righe 1, 3, 7, 9 della Figura 17.4) siano rappresentative della stessa popolazione rappresentata dalle sole unità estratte nel 2015 e sopravvissute nel 2016 (cioè quelle appartenenti al panel alla seconda intervista, riga 1 in Figura 17.4). A tal fine, si parte dai pesi già corretti per mancata risposta (RB060) disponibili per le altre unità compresenti nella coppia di anni 2015-2016 (cioè riga 3 del panel alla terza intervista e righe 7 e 9 del panel alla quarta intervista) e si riproporzionano i pesi del 2016 in modo che la somma dei pesi dei compresenti nei due anni di ciascun panel sia uguale alla somma

dei pesi dei compresenti osservati sul solo panel alla seconda intervista (P2). Si dividono i pesi così ottenuti per 3, in modo tale che la somma dei pesi per i tre panel riporti alla popolazione estratta nel 2015 e sopravvissuta nel 2016 (Figura 17.5c). Anche in questo caso, si è preferito ricorrere a un semplice riproporzionamento in quanto l'utilizzo di modelli di selezione ad-hoc fornisce stime del tutto analoghe.

Transizioni tra due, tre o quattro anni

Con la stessa logica descritta nel paragrafo precedente, si possono stimare le transizioni a distanza di due anni, sfruttando le unità campionarie compresenti in tre anni consecutivi, appartenenti non solo al panel alla terza intervista ma anche a quello alla quarta (rispettivamente righe 3 e 7 in Figura 17.4). I pesi forniti per queste unità compresenti nei tre anni (RB063) sono rappresentativi della popolazione estratta nel 2014 e sopravvissuta fino al 2016, il cui ammontare coincide con quello della popolazione longitudinale del panel alla terza intervista (Figura 17.5b).

Tra gli indicatori prodotti da Eurostat che usano questo tipo di strategia, ad esempio, ci sono le transizioni tra decili osservate tra 2, 3 o 4 anni (Tavola 17.1). Il confronto tra questi indicatori mostra come la persistenza nello stesso decile diminuisca con l'allungarsi del periodo di osservazione, ovvero coloro che non cambiano decile tra due anni consecutivi sono il 49,2 per cento della popolazione, mentre quelli che non cambiano decile in 3 o 4 anni sono rispettivamente il 43,5 per cento e il 39,7 per cento. È importante notare però che tali indicatori, pur fornendo indicazioni rilevanti, sono stimati su collettivi diversi (ovvero i compresenti in 4, 3 o 2 anni), utilizzando pesi diversi (rispettivamente RB064, RB063 e RB062) (Eurostat, 2014, pp: 269-273).

Tavola 17.1 - Transizioni tra decili, osservate tra due, tre e quattro anni

TRANSIZIONI VERSO:	Dopo 1 anno	Dopo 2 anni	Dopo 3 anni
Un solo decile più alto	16,3	17,0	16,9
Più di un decile più alto	10,3	12,8	14,4
Un solo decile più basso	13,5	13,7	15,2
Più di un decile più basso	10,7	13,1	13,8
Nessun cambiamento	49,2	43,5	39,7

17.4 Conclusioni

La struttura campionaria su cui si basa EU-SILC offre molteplici potenzialità di analisi e stima, secondo gli obbiettivi conoscitivi e di ricerca. Con un opportuno sistema di pesi, le unità di rilevazione appartenenti ai diversi panel possono essere combinate per ottenere stime sia trasversali, sia longitudinali.

Al fine di sfruttare queste potenzialità offerte dalla struttura campionaria, come da Regolamento Europeo, l'Istat stima, oltre ai pesi trasversali, un sistema di pesi longitudinali così articolato:

- i pesi che descrivono, per ciascun panel e ciascun anno, l'evoluzione nel tempo delle caratteristiche della popolazione della prima intervista (RB060). Questi pesi sono valorizzati, in ogni anno di rilevazione, per tutti i componenti familiari delle famiglie

intervistate nell'anno di estrazione del campione e rilevati con successo negli anni successivi, oltre che per i nuovi nati durante il periodo di osservazione del panel;

- i pesi che consentono di ottenere stime longitudinali sulla base delle sole unità di rilevazione effettivamente intervistate in tutti e quattro gli anni di osservazione del panel completo. Questi pesi sono rilasciati solo per l'ultimo anno di rilevazione *t* e sono rappresentativi della popolazione estratta nell'anno *t-3* e sopravvissuta fino all'anno più recente di rilevazione *t* (RB064);
- i pesi che consentono di ottenere stime su tutte le unità di rilevazione osservate negli ultimi tre anni di rilevazione e quindi appartenenti ai due panel intervistati in questo periodo. Tali pesi sono rilasciati solo per l'anno di rilevazione più recente *t*, e sono rappresentativi della popolazione estratta nell'anno *t-2* e sopravvissuta fino all'anno di rilevazione *t* (RB063);
- i pesi che consentono di avere stime sfruttando tutte le unità di rilevazione osservate negli ultimi due anni e quindi appartenenti ai tre panel intervistati in questi anni. Questi pesi, rilasciati solo per l'ultimo anno di rilevazione *t*, sono rappresentativi della popolazione estratta nell'anno *t-1* e sopravvissuta fino all'anno *t* (RB062).

Nella procedura di stima del sistema di pesi prodotti, si tiene in considerazione il processo di selezione associato alla mancata risposta, almeno secondo alcune caratteristiche individuali e familiari osservate alla prima intervista, in modo tale che le unità effettivamente intervistate siano rappresentative anche di quelle uscite dall'osservazione per *attrition* prima del quarto anno di osservazione, e quindi dell'intera popolazione di partenza.

RIFERIMENTI BIBLIOGRAFICI

Bagatta, G. (*a cura di*). 2006. "Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini". *Metodi e Norme*, n. 31. Roma: Istat.

Barcaroli, G., L. D'Aurizio, A. Manzari, e A. Pallara. 1999. "Metodi e software per il controllo e la correzione dei dati". *Documenti*, n. 1. Roma: Istat.

Bazzoli, M., S. Marzadro, A. Schizzerotto, e U. Trivellato. 2018. "Un'esperienza pilota di integrazione di dati amministrativi e di survey per l'analisi dell'evoluzione delle storie lavorative dei giovani". *FBK-IRVAPP Working Paper Series*, N. 2018-01. Trento: Research Institute for the Evaluation of Public Policies - IRVAPP.

Betti, G., G. Donatiello, and V. Verma. 2011. "The Siena Microsimulation Model (SM2) for net-gross conversion of Eu-Silc income variables". *International Journal of Microsimulation*, International Microsimulation Association, Volume 4 (1): 35-53.

Bricker, J., A.M. Henriques, J.A. Krimmel, and J.E. Sabelhaus. 2015. "Measuring Income and Wealth at the Top Using Administrative and Survey Data". *Finance and Economics Discussion Series*, 2015-030: 1-63. Washington, DC, U.S.: Board of Governors of the Federal Reserve System.

Budano, G., e S. Demofonti (*a cura di*). 2010. "La misurazione delle tipologie familiari nelle indagini di popolazione". *Metodi e Norme*, n. 46. Roma: Istat.

Burricand, C. 2013. "Transition from survey data to registers in the French SILC survey". In Jäntti, M., V.-M. Törmälehto, and E. Marlier (*eds.*). "The use of registers in the context of EU–SILC: challenges and opportunities". *Eurostat Statistical working papers*: 111-124. Luxembourg: Publications Office of the European Union.

Calderwood, L., and C. Lessof. 2009. "Enhancing longitudinal surveys by linking to administrative data". In Lynn, P. (*ed.*). *Methodology of Longitudinal Surveys*. Hoboken, NJ, U.S.: John Wiley & Sons.

Ceccarelli, C., M. Di Marco, e C. Rinaldell (*a cura di*). 2008. "L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)". *Metodi e Norme*, n. 37. Roma: Istat.

Coli, A., P. Consolini, and M. D'Orazio. 2016. "Administrative and Survey Data Collection and Integration". In Pratesi, M (*ed.*). *Analysis of Poverty Data by Small Area Estimation*, Chapter 3: 41-60. Hoboken, NJ, U.S.: John Wiley & Sons.

Consolini, P. 2000. "Le prestazioni sociali monetarie non pensionistiche: aspetti istituzionali e classificazioni statistiche". *Documenti*, n. 2. Roma: Istat.

Consolini, P. 2004. "L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie". *Contributi*, n. 29. Roma: Istat.

Consolini, P. 2015. "A Methodology for integration of survey and administrative data". In *IT-Silc project*. Presentation at the *Study Visit on IT tools for record Linkage, statistical matching and SILC survey*. Roma, Italy, 8th – 9th June 2015.

Consolini, P., and G. Donatiello. 2013. "*Improvements of data quality through the combined use of survey and administrative sources and micro simulation model*". In Jäntti, M., V.-M. Törmälehto, and E. Marlier (*eds.*). "The use of registers in the context of EU–SILC: challenges and opportunities". *Eurostat Statistical working papers*: 125-139. Luxembourg: Publications Office of the European Union.

Consolini, P., e R. De Carli. 2002. "Le prestazioni sociali monetarie non pensionistiche: unità di analisi, fonti e rappresentazione statistica dei dati". *Documenti*, n. 1. Roma: Istat.

Consolini, P., and G. Donatiello. 2015. "Multi-source data collection strategy and microsimulation techniques for the Italian EU-SILC". *Rivista di statistica ufficiale*, N. 2/2015: 77-96. Roma: Istat. https://www.istat.it/it/archivio/171133.

Consolini, P. (*a cura di*). 2009. "Integrazione dei dati campionari EU-SILC con dati di fonte amministrativa". *Metodi e norme*, n. 38. Roma: Istat.

Delle Fratte, C., and F. Lariccia. 2015. "The impact of Administrative data on final estimates of It-Silc income variables". Presentation at the *Workshop on best practices for EU-SILC revision*, London, U.K., September 2015.

Delle Fratte, C., e F. Lariccia. 2016. "L'impatto dei dati amministrativi sulle stime finali dei redditi dell'indagine IT-SILC". *Rivista Italiana di Economia, Demografia e Statistica*, Volume LXX, N. 3: 113-124.

Deville, J.-C., and C.E. Särndal. 1992. "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, Volume 87, n. 418: 376-382.

Donatiello, G. (*a cura di*). 2011. "La metodologia di stima dei redditi lordi nell'indagine Eu-Silc. Indagine europea sui redditi e le condizioni di vita delle famiglie". *Metodi e norme*, n. 49. Roma: Istat.

Donatiello, G., G. Betti, and P. Consolini. 2012. "The Construction of Gross Income Variables of Eusilc (Eu Statistics on Income and Living Conditions) in Italy: A Mixed Strategy Using Microsimulation and Administrative Data". *Quaderni del Dipartimento di Economia Politica e Statistica*, n. 652. Siena, Italy: Università degli Studi di Siena.

Dunn, H.L. 1946. "Record Linkage". *American Journal of Public Health*, Volume 36 (12): 1412–1416.

European Commission. 2014. *Method: Reconciling Conflicting Microdata*. Memobust Handbook on Methodology of Modern Business Statistics, Luxemburg.

European Commission. 2003. *Commission Regulation (EC) No 1982/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the sampling and tracing rules*.

European Commission, Eurostat. 2017. *Methodological guidelines and description of EU-SILC Target Variables (Version October 2016)*. Luxemburg: Eurostat.

European Commission, Eurostat. 2014. "Working paper with the description of Income and Living conditions dataset". *EU-SILC: Methodological studies and publication*. Luxembourg: Eurostat.

European Commission, Eurostat. 2012. *Comparative Intermediate Quality Report 2010. Version 3 - October 2012*. Luxembourg: Eurostat.

European Commission, Eurostat. 2008. *Comparative Final EU Quality Report 2005 (Version 2 - September 2008)*. Luxembourg: Eurostat.

European Commission, DGINS. 2011. Wiesbaden memorandum. New conceptual design for household and social statistics. https://circabc.europa.eu/sd/a/51a4bcbd-2ac2-46a8-8992-cfb9e6009522/Item_3.3.%20Modernisation%20of%20Social%20statistics_annex.pdf.

European Parliament and Council. 2003. *Regulation (EC) No 1177/2003 of the European Parliament and of the Council of 16 June 2003 concerning Community statistics on income and living conditions (EU-SILC)*.

Fellegi, I.P., and D. Holt. 1976. "A Systematic Approach to Automatic Edit and Imputation". *Journal of the American Statistical Association*, Volume 71, N. 353: 17-35.

Fellegi, I.P., and A. Sunter. 1969. "A Theory for Record Linkage". *Journal of the American Statistical Association*, Volume 64, Issue 328: 1183–1210.

Freguja, C., e M.C. Romano (*a cura di*). 2014. "La modernizzazione delle tecniche di rilevazione nelle indagini socio-economiche sulle famiglie". *Letture Statistiche – Metodi*. Roma: Istat. https://www.istat.it/it/archivio/145721.

Gower, J.C. 1971. "A General Coefficient of Similarity and Some of Its Properties". *Biometrics*. Volume 27, N. 4: 857-871.

Granquist, L., and J.C. Kovar. 1997. "Editing of survey data: How much is enough?". In Lyberg, L.E., P.P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (*eds.*). *Survey*

measurement and process quality: 415-435. Hoboken, NJ, U.S.: John Wiley & Sons, *Series in Probability and Statistics*.

Hoogendoorn, A.W. 2004. "A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing". *Journal of Official Statistics - JOS*, Volume 20, N. 2: 219–232.

Istituto Nazionale della Previdenza Sociale - INPS. 2021. *Il calcolo dei contributi e le aliquote (www.inps.it)*. Roma: INPS.

Istituto Nazionale della Previdenza Sociale – INPS. 2015. Rapporto annuale 2014. Roma: INPS.

Istituto Nazionale della Previdenza Sociale – INPS, Direzione Centrale Organizzazione e Sistemi Informativi - Area Gestione Aziende e Lavoratori Dipendenti. 2015. *Documento tecnico per la compilazione dei flussi delle denunce retributive e contributive individuali mensili. UNIEMENS (individuale), Release 3.2.1 del 17/12/2015*.

Istituto Nazionale di Statistica - Istat. 2015. "*Condizioni di vita (Eu-Silc). Dati trasversali*". Roma: Istat. http://www.istat.it/it/archivio/4152.

Istituto Nazionale di Statistica - Istat. 2004. "CONCORD V.1.0 – Controllo e correzione dei dati. Manuale utente e aspetti metodologici". *Tecniche e Strumenti*, N. 1/2004. Roma: Istat. https://www.istat.it/it/files/2011/03/manualeconcord.pdf.

Jäckle, A. 2008. "Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection". *Journal of Official Statistics - JOS*, Volume 24, N. 3: 411–430.

Kalton, G., D. Kasprzyk, and D.B. Mc Millen. 1989. "Non-sampling Errors in Panel Surveys". In Kasprzyk, D., G. Duncan, G. Kalton, M.P. Singh (*eds.*). *Panel Surveys*: 249-270. New York, NY, U.S.; Chichester, UK: John Wiley & Sons.

Karpinski, M., and A. Zelikovsky. 1997. "Approximating dense cases of covering problems". *Electronic Colloquium on Computational Complexity - ECCC*, Volume 4.

Korinek, A., J.A. Mistiaen, and M. Ravallion. 2006. "Survey nonresponse and the distribution of income". *The Journal of Economic Inequality*, 4: 33-55.

Lo Castro, D. 2016. *Indagine sul reddito e le condizioni di vita (EU-SILC)*. In Istituto Nazionale di Statistica - Istat., *Navigando tra le fonti sociali*. Roma: Istat. http://schedefontidati.istat.it/index.php/Navigando_tra_le_fonti_sociali.

Lynn, P., A. Jäckle, S.P. Jenkins, and E. Sala. 2006. "The Effects of Dependent Interviewing on Responses to Questions on Income Sources". *Journal of Official Statistics - JOS*, Volume 22, N. 3: 357–384.

Lund, C., and M. Yannakakis. 1994. "On the hardness of approximating minimization problems". *Journal of the ACM – Association of Computer Machinery*, Volume 41, N. 5: 960–981.

Massoli, P. 2008. "Using Graphs Theory for Modelling a Survey Questionnaire". Paper presented at the *European Conference on Quality in Official Statistics – Q2008*. Roma, Italy, 8th – 11th July 2008.

Ministero dell'Economia e delle Finanze – MEF, Dipartimento delle Finanze. 2016. *Dichiarazioni fiscali*. Roma: MEF.

Pannekoek, J. 2011. "Models and algorithms for micro-integration". In Istat, CBS, GUS, INE, SSB, SFSO, Eurostat. *Report on WP2: Methodological developments. ESSnet on Data Integration*. Chapter 6: 120-131. Luxembourg: Eurostat. https://ec.europa.eu/eurostat/cros/system/files/WP2.pdf.

Spinelli, V., e M. Tancioni. 2004. "I Trattamenti Monetari non Pensionistici: Approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10. Anni 1999-2001". *Contributi*, n. 28. Roma: Istat.

Trivellato, U. 2017. "Microdata for social sciences and policy evaluation as a public good". *FBK-IRVAPP Working Paper Series*, N. 2017-06. Trento: Research Institute for the Evaluation of Public Policies - IRVAPP.

United Nations Economic Commission For Europe - UNECE. 2011. *Canberra Group Handbook on Household Income Statistics. 2nd Edition 2011*. New York, NY, U.S.; Geneva, Switzerland: United Nations.

van der Laan, P. 2000. "Integrating administrative registers and household surveys". *Summer 2000 Special Issue*, Volume 15. Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

Wallgren, A., and B. Wallgren. 2007. *Register-based Statistics: Administrative Data for Statistical Purposes*. Chichester, UK: John Wiley & Sons Ltd, *Series in Survey Metodology*.