ISFOL

L'indagine campionaria Isfol-PLUS: contenuti metodologici e implementazione

Michele Giammatteo

ISSN 1974-4978

collana Studi Isfol | numero 2009/3- marzo | www.isfol.it | tema occupazione

L'Istituto per lo sviluppo della formazione professionale dei lavoratori (Isfol) è un ente pubblico istituito con DPR n. 478 del 30 giugno 1973. Nasce per accompagnare la prima fase di decentramento regionale delle competenze in materia di formazione professionale, codificata nella legge n. 845 del dicembre 1978; dal 1999 viene incluso tra gli enti pubblici di ricerca con DL n. 419 del 29/10/1999. L'attuale Statuto, approvato con DPCM del 19 marzo 2003, sancisce per l'Istituto competenze nel campo delle politiche formative, del lavoro e sociali.
L'Isfol svolge e promuove attività di studio, ricerca, sperimentazione, documentazione, valutazione, informazione, consulenza e assistenza tecnica per lo sviluppo della formazione professionale, delle politiche sociali e del lavoro. Contribuisce al miglioramento delle risorse umane, alla crescita dell'occupazione, all'inclusione sociale e allo sviluppo sociale. È sottoposto alla vigilanza del Ministero del lavoro e della previdenza sociale al quale fornisce supporto tecnico-scientifico ed opera in collaborazione con il Ministero della pubblica istruzione, il Ministero della solidarietà sociale, la Presidenza del Consiglio dei ministri, le Regioni, le Parti sociali, l'Unione europea e altri Organismi internazionali.

Studi Isfol, la prima collana scientifica elettronica realizzata dall'Isfol, comprende articoli e *working paper* sui temi della formazione, del lavoro, dell'inclusione sociale.
La collana nasce con l'intento di rendere accessibili a tutti liberamente, idee e dati, anche nel corso della loro elaborazione. In particolare, mira a stimolare il dibattito e la circolarità delle riflessioni nella comunità scientifica, offrendo l'opportunità, grazie alla sua multimedialità, di creare intorno ad essi una *community*.

La Collana *Studi Isfol* è curata da *Claudio Bens*i – Responsabile Servizio comunicazione web e multimediale
Coordinamento editoriale: *Paola Piras, Aurelia Tirelli, Matilde Tobia*
Progetto grafico: *Marco Boccia*
Contatti: editoriadigitale@isfol.it

La presente pubblicazione costituisce la versione cartacea dell'edizione consultabile sul portale *www.isfol.it* all'interno della collana elettronica *Studi Isfol*.

Indice

Gli autori

Michele Giammatteo
Ricercatore Isfol

1. Introduzione[1]

L'indagine Isfol-PLUS è una rilevazione campionaria alla sua seconda annualità[2]. Essa, presente nel Piano Statistico Nazionale, ha l'obiettivo di mettere a disposizione delle istituzioni e dei ricercatori la base informativa necessaria alle analisi del mercato del lavoro italiano, in un'ottica di complementarietà con le fonti nazionali già disponibili (Istat e Inps) e secondo un approccio integrato sulla base di due principali filoni di ricerca, economico e sociologico.

L'indagine è stata eseguita nel II-III trimestre del 2006 ed ha coinvolto 37.513 individui[3]. Questi sono stati contattati attraverso interviste telefoniche – effettuate con sistema CATI – somministrate esclusivamente al rispondente[4], facenti riferimento a condizioni individuali a carattere oggettivo e riguardanti: le dinamiche occupazionali dei giovani, delle donne e delle persone con più di 50 anni, della disoccupazione, dei percorsi formativi e della ricerca di lavoro.

Il processo di raccolta dati è stato affidato per il secondo anno consecutivo (come previsto da bando) alla società di rilevazione Doxa Spa. La seconda annualità di rilevazione con la stessa società ha permesso di minimizzare gli attriti di natura procedurale, dalla pianificazione alla fornitura finale della banca dati, che in genere possono sorgere in caso di cambio di impresa appaltatrice. È da sottolineare, inoltre, la maggiore complessità dell'indagine PLUS 2006 rispetto all'esperienza dell'anno precedente, determinata soprattutto dalla consistente componente *panel* del campione. Oltre il 62% degli individui intervistati nel 2005 sono stati reintervistati ad un anno di distanza[5] dal primo contatto, implicando una disponibilità di informazione longitudinale in grado di arricchire ulteriormente le potenzialità di analisi dei dati raccolti.

Prima di approfondire nel dettaglio gli aspetti relativi al piano di campionamento, alla metodologia di definizione dei pesi *cross-section* e *panel* di riporto all'universo e all'imputazione delle mancate risposte parziali, vogliamo di seguito evidenziare le principali caratteristiche dell'indagine PLUS, mettendone in evidenza le potenzialità, il rigore concettuale e metodologico alla base della sua

[1] Un ringraziamento particolare a Piero D. Falorsi (ISTAT) per il supporto scientifico fornito in molte delle fasi di lavoro descritte nel presente studio; a Gianni Corsetti (ISFOL) per il suo contributo ai paragrafi 5 e 6; ad Alessandro Martini (ISFOL) per la preparazione del paragrafo 5.1.

[2] La terza annualità, PLUS 2008, le prime elaborazioni e il file standard utile ai fini della comunicazione esterna all'Isfol, saranno disponibili entro l'estate 2009.

[3] Nel 2005 la rilevazione ha raccolto poco più di 40.000 interviste. La differenza campionaria tra i due anni è dovuta principalmente a due fattori: l'esclusione dal campione 2006 delle donne casalinghe tra 40 e 49 anni e la necessità di limitare l'arco temporale di rilevazione, resa più onerosa dalla necessità di effettuare una consistente quota di interviste *panel*.

[4] Soltanto marginalmente le domande incluse nel questionario hanno previsto la possibilità di raccogliere informazioni anche sugli altri componenti del nucleo familiare dell'intervistato (informazioni di tipo *proxy*), quali: la numerosità del nucleo familiare, la presenza di under 15 o over 65, ecc.

[5] Fino ad un massimo di 15 mesi.

progettazione - prerequisito imprescindibile dell'analisi dell'informazione raccolta - anche proponendo un confronto con la maggiore fonte di informazione statistica nazionale, costituita dalla Rilevazione continua sulle forze di lavoro (RCFL) dell'Istat.

Il primo aspetto che ci sembra importante affrontare riguarda la definizione di alcuni aggregati caratterizzanti le analisi standard sul mercato del lavoro. I dati PLUS sull'occupazione sono ispirati ad un criterio classificatorio lievemente diverso rispetto ai dati RCFL. Infatti, mentre la rilevazione PLUS definisce come occupati e in cerca di lavoro le persone che si *auto definiscono* tali, RCFL segue un percorso che identifica la condizione in base ad alcune informazioni "oggettive" che riguardano: per gli occupati l'aver lavorato almeno un'ora nella settimana di riferimento dell'intervista e per le persone in cerca l'aver compiuto almeno un tentativo di ricerca, ed essere immediatamente disponibili a lavorare. Da questa dovuta precisazione ne segue come l'impianto Istat (Eurostat) sottenda nel proprio meccanismo contatore una certa inclinazione a considerare gli individui più facilmente occupati e meno persone in cerca.

L'idea generale dell'indagine PLUS di registrare, nel modo più accurato possibile, la condizione *auto percepita* dai soggetti intervistati fa sì che anche la distinzione tra *persone in cerca* ed *inattivi* sia differente da quanto adottato in RCFL.

In particolare:

a) si considerano *persone in cerca*, e quindi attive, alcune tipologie di individui che per l'Istat sono da considerare *inattivi*

b) non si considerano occupati quei soggetti che svolgono una attività lavorativa che non è, in termini economici e secondo la propria percezione, tale da giustificare la loro inclusione in tale categoria (studenti, pensionati da lavoro e casalinghe – lavoratrici/ori saltuari), considerandoli *occupati non prevalenti*.

Perché questa scelta? Essendo il fine di PLUS quello di verificare la qualità dell'attuale occupazione è stato necessario riferirsi allo status percepito dall'intervistato, poiché utilizzando uno schema analogo a quello dell'Istat, ovvero seguendo le indicazioni dei regolamenti comunitari, si correva il rischio di includere tra gli *occupati* quelli *con condizione non prevalente* e i *disoccupati non attivi*, che rappresentano, invece, le sottopopolazioni di maggior interesse per l'attivazione stabile e continuativa.

Ovviamente da PLUS è possibile ricostruire gli occupati nelle definizioni Istat-Eurostat essendo stati somministrati i quesiti necessari alla loro individuazione; pertanto abbiamo deciso di vincolare i dati PLUS ad alcuni aggregati ufficiali di fonte RCFL, precisando che gli *occupati non prevalenti* e gli inattivi che si dichiarano in cerca sono stati considerati nelle condizioni cui si attribuivano autonomamente.

La prima conseguenza è che le caratteristiche dell'occupazione sono al netto della componente *occupati non prevalenti* (ovvero gli individui considerati occupati secondo la definizione Istat ma che abbiano un'attività economicamente non tale da farli annoverare tra gli occupati *tout court*) e al lordo di alcune categorie di disoccupati (o in cerca di lavoro) che invece non rientrano nella definizione Istat (*inattivi che cercano lavoro*). Questo rende possibile la somministrazione di quesiti estremamente dettagliati sulla natura e le caratteristiche del lavoro, dell'istruzione e della condizione familiare, consentendo – attraverso moduli dedicati – di fornire stime attendibili anche per aggregati molto piccoli o temi specifici[6]. L'impianto PLUS consente stime di aggregati anche relativamente poco numerosi nella popolazione (70.000-100.000 individui), pari a circa lo 0,5% dell'occupazione, con una probabilità del 95% che l'intervallo compreso tra ±5% del valore stimato comprenda il valore vero corrispondente[7].

Nella tabella 1 è riportato un confronto tra le stime RCFL e PLUS. Premettiamo che secondo le classificazioni Istat si intendono per collaboratori i soli *collaboratori puri*, cioè coloro che pur essendo lavoratori autonomi non sono professionisti e svolgono la loro attività prevalentemente attraverso forme di lavoro di collaborazione coordinata e continuativa o a progetto. L'indagine PLUS identifica anche i finti collaboratori (definiti *parasubordinati*), ovvero quegli occupati con forme di lavoro autonome che svolgono lavori con modalità tipiche del lavoro dipendente. Una voce a sé stante è stata prevista per le finte Partite IVA e le collaborazioni occasionali, che sono risultate essere delle forme di lavoro fortemente subordinate.

Si rammenta che il confronto tra le due popolazioni complessive è possibile soltanto prendendo in considerazione la sottopopolazione RCFL (2006) definita dai domini di studio PLUS[8]. La scelta di focalizzare l'attenzione su una particolare – seppure molto significativa – quota della popolazione italiana è coerente con l'obiettivo prioritario dell'indagine PLUS: fornire stime attendibili di fenomeni rari e marginali, ovvero analizzare nel dettaglio le composizioni degli stock di riferimento forniti dall'Istat.

[6] In particolare, i fenomeni caratterizzanti l'atipicità dell'occupazione, la qualità e la ricerca del lavoro, le caratteristiche dei giovani, delle donne, degli over 50, la condizione delle persone in cerca di lavoro e degli inattivi.

[7] Vedi i paragrafi 2 e 5 per una descrizione dettagliata dei criteri e metodologie utilizzate nel piano di campionamento e analisi della varianza delle stime.

[8] Vedi la definizione dei target di indagine nel successivo paragrafo. Sono, ad esempio, esclusi gli studenti maschi oltre i 30 anni e le studentesse oltre i 39, i pensionati al di sotto dei 50 anni, gli over 64 e under 15 in tutte le condizioni.

Tabella 1 - Confronto tra le stime Istat-RCFL e Isfol-PLUS

Istat - RCFL		Isfol - PLUS	
Definizione	**Numerosità**	**Numerosità**	**Definizione**
Dipendenti permanenti	14.638.756	14.253.628	Dipendente a tempo indeterminato
		423.798	Altro Dipendente**
Dipendenti a termine	2.212.998	1.075.122	Dipendente a tempo determinato
		1.099.186	Altre forme lavoro dipendente a termine*
Totale Dipendenti	16.851.755	16.851.733	
Collaboratori	478.911	713.637	Colla. coord. e continuative e a progetto
Collaboratori e consulenti		579.761	P. IVA e Collaborazioni Occasionali
Autonomi	5.287.864	4.281.682	Imprenditori e professionisti
		191.699	Altro Autonomo**
Totale Autonomi	5.766.775	5.766.779	
Totale Occupati	22.618.530	22.618.512	

Fonte: Elaborazioni su microdati Istat-RCFL e Isfol-PLUS 2006

* CFL, Apprendistato, Contratto d'inserimento, Lavoro interinale o a somministrazione, Job sharing o lavoro ripartito, Lavoro intermittente o a chiamata, Alternanza scuola-lavoro, Stage, Pratica professionale, Tirocinio.
** Lavoro con "contratto informale" - "non sa o non ricorda": riclassificati in lavoro dipendente o autonomo attraverso l'utilizzo di controlli previsti nel questionario e, per una componente residua, su base probabilistica.

2. Il piano di campionamento

Il disegno dell'indagine PLUS 2006 non ha subito variazioni significative rispetto alla procedura seguita nel 2005. La rilevazione dei dati tramite interviste CATI senza rispondenti *proxy* è stata effettuata sulla base di un *campionamento stratificato per quote* con definizione di domini di studio parzialmente sovrapposti.

Il campione è stato suddiviso negli stessi cinque target fondamentali del 2005 (o domini di interesse) costituiti da: i giovani tra i 15 e i 29 anni, le donne tra 20 e 39 anni[9], la popolazione in età compresa tra 50 e 64 anni, le persone non occupate in cerca di lavoro e gli occupati tra i 15 e i 64 anni. Questi sono stati opportunamente disaggregati per classe d'età e condizione, in modo da ottenere i seguenti 9 domini di studio:

1. *Giovani occupati*, in età compresa tra 15 e 29 anni
2. *Giovani studenti*, in età compresa tra 15 e 29 anni
3. *Giovani altra condizione*, in età compresa tra 15 e 29 anni
4. *Donne attive*, in età compresa tra 20 e 39 anni

[9] Questa rappresenta l'unica variazione rispetto a PLUS 2005, che comprendeva anche le donne casalinghe tra 40 e 49 anni.

5. *Donne inattive*, in età compresa tra 20 e 39 anni
6. *Anziani attivi*, in età compresa tra 50 e 64 anni
7. *Anziani inattivi* (pensionati da lavoro), in età compresa tra 50 e 64 anni
8. *In cerca* (definizione estesa)[10]
9. *Occupati*.

Allo scopo di poter fornire stime attendibili anche per sottopopolazioni di questi 9 domini si è proceduto alla pianificazione di un *campionamento stratificato*, dove gli strati – definiti dall'incrocio delle variabili riportate in tabella 2 – costituiscono una partizione del campione e (per sottoinsiemi di strati) degli stessi domini di studio. Il numero di interviste da effettuare per ciascuno degli strati è stato determinato in modo da fornire stime attendibili per l'intera popolazione di riferimento e per particolari sottoinsiemi d'interesse[11].

Formalmente possiamo rappresentare la numerosità del generico dominio *d* nella popolazione come,

$$N_d = \sum_{s=1}^{H} N_s I_{s,d}$$

dove N_s è la numerosità del generico strato *s* nella popolazione e $I_{s,d}$ è una variabile indicatrice che vale 1 se lo strato *s* è contenuto in *d* e 0 altrimenti.
Rimandando al capitolo 9 del Rapporto PLUS 2005[12] il dettaglio della trattazione teorica della metodologia di allocazione campionaria negli strati, è sufficiente qui ricordare l'obiettivo finale della procedura: definire, per ciascuno degli strati *s*, una numerosità campionaria n_s tale da garantire che la generica stima p_d della proporzione P_d nella popolazione abbia una varianza minima. Ciò significa che l'incidenza di un qualsiasi fenomeno per un sottocollettivo di individui può essere efficacemente stimato dai dati con un errore molto piccolo, o meglio, statisticamente non significativo.
Questo problema è tanto intuitivo quanto complesso dal punto di vista computazionale. Basti pensare che esso è da estendersi a *tutti* gli strati $s=1,\ldots,H$, che possono contenere individui appartenenti contemporaneamente a più di un dominio di studio[13] - allocazione *multidomain* - e che si deve ottenere una ed una sola soluzione come risultato della contemporanea soluzione di *H* problemi di minimizzazione vincolata. Il tutto viene risolto attraverso l'implementazione di un algoritmo di iterazione che converge ad un unico risultato ottimo ($n^*_1, \ldots, n^*_H$), attraverso

[10] Vedi tabella 3.
[11] Per le 20 regioni e per i 13 comuni italiani con popolazione superiore a 250.000 abitanti, disaggregati a loro volta per genere, classi di età e condizione occupazionale.
[12] Mandrone E., Radicchia D. (a cura di), *PLUS - Participation Labour Unemployment Survey*, Roma, Isfol, 2006 (I libri del Fondo sociale europeo).
[13] Si pensi, ad esempio, ad un occupato di 25 anni maschio che appartiene sia al dominio 1 che 9.

la risoluzione di un sistema di allocazione campionaria negli strati vincolata a livelli di varianza predefiniti.

Il campione dell'indagine PLUS rientra nella categoria dei campioni "non probabilistici", più in particolare esso è un *campione stratificato per quote*. Ricadono sotto la denominazione di campioni non probabilistici quelli in cui la probabilità di inclusione delle unità di rilevazione non è conosciuta *a priori*, a causa dell'assenza di liste dalle quali selezionare gli individui da intervistare. Per ovviare a questo problema, la popolazione di riferimento è ricavata dalle stime prodotte dalla RCFL dell'ISTAT, la quale fornisce le informazioni utili alla determinazione delle quote campionarie e quindi delle interviste da realizzare per ciascuno degli strati predefiniti[14].

Tabella 2 – Variabili di stratificazione del campione PLUS 2006

Variabili	Modalità
Regione	Piemonte e Valle d'Aosta, Lombardia, Trentino A.A., Veneto, Friuli V.G., Liguria, Emilia Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna
Tipo comune	Comune metropolitano, Comune non metropolitano
Sesso	Maschi, Femmine
Età in classi	15-19, 20-29, 30-39, 40-49, 50-64
Condizione occupazionale	Occupato, In cerca di occupazione, Studente, Pensionato da lavoro, Altro inattivo (casalinga)

Fonte: Isfol PLUS 2006

Come accennato nel paragrafo precedente, la seconda annualità dell'indagine PLUS ha previsto la reintervista di una quota consistente di individui già contattati nella prima *wave*. La scelta operata al momento della rilevazione è stata quella di limitare in modo non stringente le interviste *panel*, attraverso l'imposizione di due semplici vincoli:

a) riempimento massimo degli strati con interviste *panel* al 70%

b) una quota *panel* complessiva non superiore ai 2/3 del campione PLUS 2006.

Questa decisione è stata motivata dall'obiettivo di massimizzare la dimensione del campione longitudinale e, contemporaneamente, dalla convinzione di essere in grado di ribilanciare *ex post* lo stesso con una quota consistente di nuove interviste (*non panel*). In questo modo si sono potute efficacemente garantire sia le stime prodotte in termini sezionali che quelle di ambito longitudinale, affidando

[14] Per un approfondimento sulle strategie di campionamento, allocazione campionaria e relative tecniche di stima si veda Centra M. e Falorsi P.D. (2008).

poi alla fase di riporto all'universo e post-stratificazione il compito di correggere l'autoselezione del campione[15].

3. Riporto all'universo e stimatore di calibrazione sezionale

3.1. Determinazione del peso base

Nella fase successiva alla rilevazione assume un ruolo fondamentale la scelta del tipo di stimatore di ponderazione vincolata, utile ai fini del calcolo del *coefficiente di riporto all'universo*. Sulla base di un campione di limitata dimensione questo ci permette di produrre delle stime di importanti fenomeni socio-economici riferibili all'intera popolazione oggetto di studio e, cosa fondamentale, con un'elevata affidabilità statistica.
Essendo l'indagine PLUS progettata indipendentemente dalla conoscenza a priori sulla probabilità di inclusione nel campione di ciascun individuo appartenente all'universo di riferimento, non è stato possibile costruire lo stimatore secondo l'usuale disegno di Horvitz-Thompson[16]. Come per l'indagine PLUS 2005 si è reso necessario seguire un *approccio* di tipo *predittivo* basato su modelli di superpopolazione[17], ossia sulla conoscenza di alcuni totali noti desumibili dalla popolazione di riferimento (RCFL 2006 dell'Istat) rispetto ai quali vincolare le stime prodotte dall'indagine PLUS. In particolare è stato definito uno *stimatore di regressione basato su variabili strumentali*, il quale garantisce che le stime delle frequenze assolute delle variabili ausiliarie utilizzate come regressori corrispondano ai totali noti imposti. Ciò permette di calibrare la *popolazione stimata* sulla composizione demografica e occupazionale della *popolazione reale* e correggere eventuali distorsioni causate anche da fattori connessi alla fase di rilevazione, come l'autoselezione del campione dovuta alla maggiore propensione media alla risposta telefonica di talune categorie di soggetti[18]. Rimandando al paragrafo 9.3 del Rapporto PLUS 2005[19] il dettaglio dell'illustrazione formale di tale metodologia, di seguito sono descritti nel dettaglio i passi seguiti nella preparazione delle banche dati, la scelta delle variabili strumentali utilizzate e le differenze apportate rispetto alla procedura utilizzata nel 2005[20].

---

15 Vedi la descrizione delle procedure di calibrazione cross-section e longitudinale dei paragrafi 3 e 4.
16 Vedi Horvitz D.G. e Thompson D.J. (1952).
17 Vedi Dorfman A.H., Royall R.M., Valliant R. (2000).
18 Per una dettagliata descrizione degli interventi rivolti alla correzione a posteriori dell'autoselezione campionaria vedi i paragrafi 3.2 e 4.2.
19 Mandrone E., Radicchia D. (a cura di), *PLUS - Participation Labour* … (op. cit.).
20 Il peso sezionale di riporto all'universo di PLUS 2005 è stato rivisto sulla base delle modifiche apportate nel 2006. Ciò potrebbe comportare la non corrispondenza esatta di alcune stime 2005 ottenute con il nuovo file standard che verrà presto messo a disposizione degli utenti che ne faranno richiesta. Nonostante questo, è stato ritenuto opportuno apportare tali modifiche al fine di garantire una piena coerenza tra le due annualità dell'indagine e un più elevato livello di affidabilità nelle analisi di evoluzione temporale dei fenomeni.

La prima operazione effettuata è stata la selezione della popolazione Istat-RCFL 2006 (media annuale) che costituisce l'universo di riferimento per PLUS[21], escludendo gli individui al di sotto dei 15 anni e quelli al di sopra dei 64, oltre a tutte quelle categorie (come gli studenti over 39, pensionati under 50, ecc.) non incluse tra i domini di studio elencati nel precedente paragrafo. La condizione di riferimento per la definizione degli strati è stata costruita, a partire dalle informazioni contenute in RCFL come indicato nella tabella 3, mentre nella categoria comune metropolitano sono stati inclusi i comuni di tabella 4.

Tabella 3 – Derivazione della condizione PLUS dalle definizioni RCFL

Condizione PLUS	Condizione Istat-RCFL
Occupato	• occupati
In cerca di lavoro	• persone in cerca, con precedenti esperienze, ex-occupati • persone in cerca, con precedenti esperienze, ex - inattivi • persone in cerca, senza precedenti esperienze • inattivi in età lavorativa, che *cercano* non attivamente ma disponibili • inattivi in età lavorativa, che *cercano* attivamente ma non disponibili
Studente	• inattivi & "condizione occupazionale dichiarata" = Studente
Pensionato da lavoro	• inattivi & "condizione occupazionale dichiarata" = Ritirato dal lavoro
Altro inattivo	• inattivi & "condizione occupazionale dichiarata" <> Studente e Ritirato dal lavoro *ovvero* • "inattivi in età lavorativa, *non cercano* ma disponibili" • "inattivi in età lavorativa, *non cercano* e non disponibili"

Fonte: Isfol PLUS 2006 e RCFL 2006

Il passo successivo è stato quello di identificare sul database PLUS 2006 gli strati corrispondenti e definire, per ognuno di questi, il peso base $w_{0,s}$ come rapporto tra la numerosità della popolazione (RCFL) e la numerosità campionaria, in formule:

$$w_{0,s} = \frac{N_s}{n_s}$$

[21] La differenza in termini di popolazione rispetto a RCFL è di 4.722.378 individui (al netto della popolazione al di sotto dei 15 anni e al di sopra dei 64).

Tabella 4 - Comuni metropolitani inclusi in PLUS per regione

Regione	Comuni metropolitani
Piemonte / Valle d'Aosta	Torino
Lombardia	Milano
Veneto	Venezia, Verona
Liguria	Genova
Emilia Romagna	Bologna
Toscana	Firenze
Lazio	Roma
Campania	Napoli
Puglia	Bari
Sicilia	Palermo, Catania
Sardegna	Cagliari

Fonte: Isfol PLUS 2006

3.2. Calibrazione e peso sezionale finale

In questo paragrafo è descritta la procedura utilizzata per determinare il correttore individuale di calibrazione γ_k, cioè un fattore moltiplicativo che ci permette, a partire dal peso diretto illustrato dalla , di ricavare il peso finale di *riporto all'universo e post-stratificazione*:

$$w_{1,k} = w_{0,k} \cdot \gamma_k$$

Il peso definito dalla coincide con quello ottenibile attraverso l'applicazione di un modello di regressione con utilizzo di variabili strumentali su una generica caratteristica *Y* della popolazione e opportune variabili indipendenti (dette *ausiliarie*) ***x***. Nel dettaglio, dato uno strato *s* e indicando con *k* una generica unità contenuta in esso, il vettore di variabili strumentali z_k è dato dal prodotto tra il corrispondente vettore di variabili ausiliarie x_k e il peso $w_{0,k}$, in formule

$$z_k = x_k \cdot w_{0,k} \qquad \text{per} \quad k \in s$$

Grazie alle proprietà possedute dallo stimatore, tale approccio permette di ricavare per costruzione il parametro di correzione γ_k, come

$$\gamma_k = 1 + \sum_{\bar{s}} x_k \left(\sum_{s} x_k z_k \right)^{-1} z_k$$

dove i pedici delle sommatorie *s* ed $\bar{s}$ si riferiscono rispettivamente alla quota della popolazione

osservata attraverso il campione e la parte residua non osservata. Il valore di y_k così ottenuto è tale per cui è sempre soddisfatta la condizione

$$X_{REG} = \sum_s x_k w_{1,k} = X_U$$

La garantisce che i totali X_U osservabili nella popolazione di riferimento rispetto ad una delle variabili ausiliarie x siano coerenti (ovvero, coincidano) con la stima campionaria - pesata attraverso $w_{1,k}$ - della stessa variabile. Ricordiamo anche che il correttore moltiplicativo del peso base che scaturisce dalla procedura di calibrazione è tale da conservare la validità dei vincoli imposti nel calcolo dei coefficienti di riporto all'universo.

La procedura di calibrazione prevede l'identificazione, nella popolazione di riferimento individuata su base RCFL, di un certo numero di totali noti a cui vincolare le stime pesate ottenute dai microdati PLUS. Oltre a concentrare la scelta su variabili ausiliarie correlate con le variabili di studio, si è anche tenuto conto della necessità di garantire la coerenza con l'informazione ufficiale dell'Istat in due ambiti fondamentali:

a) la distribuzione territoriale e composizione socio-demografica della popolazione
b) le caratteristiche necessarie alla definizione degli indicatori standard del mercato del lavoro (popolazione di occupati, disoccupati e inattivi – studenti, casalinghe e pensionati da lavoro).

La tabella 5 sintetizza le variabili (e rispettive modalità) prese in considerazione per la definizione dei totali noti (o *vincoli di calibrazione*).
Complessivamente sono stati imposti 142 vincoli, di cui 16 ottenuti associando alle sole unità *panel* incluse nel campione 2006 i valori di alcune variabili strutturali rilevate per quegli stessi individui nella precedente *wave*. Questo ha permesso che il peso finale sezionale di PLUS 2006 tenesse conto della distribuzione per condizione, sesso e classe d'età dei soggetti appartenenti alla popolazione longitudinale[22], incrementando la capacità di controllo dell'autoselezione del campione (*selection bias*) e garantendo, in termini più generali, una maggiore corrispondenza tra stime PLUS e RCFL.
Il punto fondamentale è dato dalla presenza di una quota n_p / n di soggetti *panel* nel campione 2006. Non avendo previsto (ed essendo difficilmente attuabile[23]) una pianificazione *a priori* delle interviste *panel* da effettuare in ogni strato, il campione sezionale pervenuto alla fine della fase di

[22] Oltretutto coerenti con i corrispondenti totali PLUS 2005.
[23] A causa della tecnica di rilevazione adottata (CATI), bassa numerosità teorica di alcuni strati non facilmente disaggregabile nelle due quote panel e non panel, ecc.

rilevazione era certamente affetto da autoselezione. Questo è dovuto sia a classici fenomeni socio-demografici, quali la mobilità territoriale, i decessi o le disgiunzioni familiari, che alla "naturale" eterogeneità della propensione alla risposta degli individui (le donne, i giovani, e le persone con titolo di studio più elevato sono risultati essere mediamente più disponibili a rilasciare una seconda intervista telefonica).

Tabella 5 – Variabili utilizzate per la definizione dei vincoli di calibrazione sezionali, PLUS 2006

Variabile 1	Variabile 2	Num. vincoli
Tipo lavoro 1 (*Dipendente, Autonomo*)	• Sesso (*M, F*) • Classe d'età (*15-19, 20-29, 30-39, 40-49, 50-64*) • Ripartizione geografica (*Nord Ovest, Nord Est, Centro, Sud Isole*)	22
Sesso	• Classe d'età • Ripartizione geografica	18
Classe d'età	• Ripartizione geografica	20
Tipo lavoro 2 (*Full time, Part time*)*	• Sesso • Classe d'età • Ripartizione geografica	22
Condizione 3 (*Occupato, In cerca, Inattivo*)	• Istruzione (*Elementare, Media, Superiore, Laurea*) • Ripartizione geografica • Sesso • Classe d'età	44 (su 45)[24]
Condizione 3 (2005) (*Occupato, In cerca, Inattivo*)	• Tipo lavoro 1 (2005) • Sesso & Classe d'età[25] (2005)	16

Fonte: Isfol PLUS 2006
* Limitata ai lavoratori dipendenti.

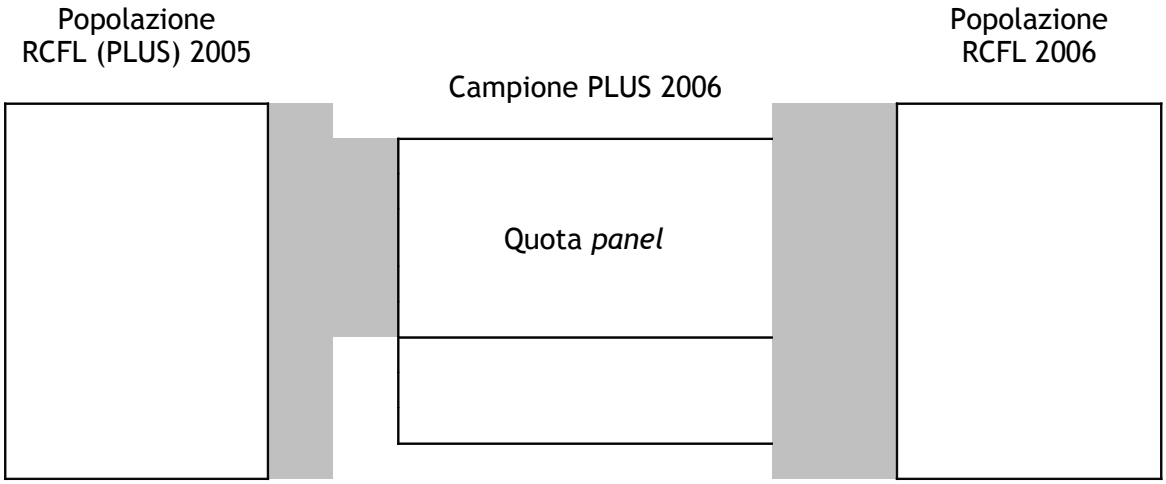
Quanto detto sopra è stato implementato attraverso l'opportuna definizione di un *peso base intermedio* ($\overline{w}_0$) di input per la procedura di calibrazione finale, costruito sulla base di due pesi *intermedi di calibrazione* ($w_{1,p}, w_{1,np}$). Questi sono stati ottenuti separatamente per la quota *panel* e per quella *non panel* attraverso due distinte procedure con, rispettivamente, 142 e 126 vincoli (vedi tabella 5). Il peso $\overline{w}_0$ è stato definito come funzione lineare dei due pesi parziali di calibrazione, con coefficienti pari all'incidenza proporzionale delle due quote (*panel* e *non panel*) sul totale del campione, in formule:

$$\begin{cases} \overline{w}_0 = w_{1,p} \cdot \dfrac{n_p}{n} & \text{se l'unità è panel} \\ \overline{w}_0 = w_{1,np} \cdot \left(1 - \dfrac{n_p}{n}\right) & \text{se l'unità è non panel} \end{cases}$$

[24] È escluso il collettivo degli inattivi 40-49 anni, non appartenente alla popolazione d'indagine.
[25] È stata costruita una variabile data dall'incrocio tra il sesso e due classi d'età: 15-39 anni e 40-64 anni.

dove $w_{1,p}$ e $w_{1,np}$ sono, rispettivamente, i pesi di output delle due calibrazione intermedie (*panel* e *non panel*), mentre n e n_p le numerosità campionarie totali e della quota *panel*. Il grafico 1 illustra le relazioni tra campione PLUS 2006, la sua quota *panel* e le popolazioni di riferimento ottenute dalla base di dati RCFL.

Grafico 1 – Schema delle relazioni esistenti tra il campione PLUS 2006 e le popolazioni di riferimento RCFL 2005 e 2006



4. Riporto all'universo e stimatore di calibrazione longitudinale

Questa sezione illustra la procedura utilizzata per la definizione del coefficiente di riporto all'universo per il *panel* PLUS 2005-2006.

Vogliamo ricordare che la disponibilità di informazione longitudinale sul mercato del lavoro disponibile in Italia è quasi completamente assorbita da due principali fonti: la rilevazione campionaria RCFL dell'Istat, e il *panel* lavoratori-imprese di fonte amministrativa Inps, sviluppato dall'Isfol e dal laboratorio Riccardo Revelli (WHIP)[26]. La prima delle due fonti rappresenta il *benchmark* italiano in termini di analisi di flusso sull'occupazione, sulla disoccupazione di media e lunga durata, ecc. La seconda costituisce un ottimo strumento di analisi di particolari fenomeni, quali: la mobilità occupazionale, quella salariale e l'analisi dei differenziali di reddito, con l'aggiunta della disponibilità di informazione congiunta su caratteristiche dell'offerta e della domanda di lavoro. L'indagine PLUS si pone come una nuova e importante fonte di analisi per alcuni fenomeni longitudinali poco approfonditi dalle basi di dati sopra ricordate. Ne costituiscono degli esempi importanti le stime delle transizioni dei lavoratori tra le diverse condizioni occupazionali e

[26] A queste si aggiungono l'indagine SHIW della Banca d'Italia e l'indagine EU-Silc coordinata in ambito EUROSTAT, ma non specificatamente fonti di informazione statistica sul mercato del lavoro.

tipologie contrattuali, o l'evoluzione nel tempo delle scelte lavorative di donne, giovani e over 50.
Al fine di costruire un *panel* di osservazioni adatto a tale scopo la prima scelta effettuata è stata la definizione di *popolazione longitudinale*. Le soluzioni possibili sono molte e variano a seconda degli obiettivi che ci si prefigge di raggiungere[27]. La scelta fatta in ambito PLUS è stata quella di considerare come riferimento la popolazione al tempo iniziale (t_0 = 2005), ossia di inserire l'analisi longitudinale nel contesto di un approccio di tipo *prospettico*[28]. In particolare, si è scelto di vincolare lo studio della composizione della mobilità occupazionale italiana alle stime prodotte da PLUS 2005 (coerenti con RCFL 2005) applicando delle *correzioni per mancata risposta panel* (attrito) e sulla base dello studio di alcuni *flussi tra condizioni di interesse primario per l'indagine*.

La dimensione finale del *panel* PLUS 2005-2006 è risultata pari al 65% del campione 2006, per una numerosità di 24.621 interviste individuali. Come già ricordato è stato riscontrato uno sbilanciamento delle interviste verso individui maggiormente istruiti, più giovani, e presumibilmente più sensibili alle tematiche che ruotano intorno al mercato del lavoro (disoccupati, occupati non permanenti, ecc.). Nelle indagini telefoniche possono in genere intervenire ulteriori fonti di distorsione nella definizione del "vero" modello di mancata risposta, quali: la limitata presenza nel domicilio (giovani *vs* meno giovani), o la rinuncia al telefono fisso.
L'obiettivo che è stato perseguito è stato quello di utilizzare tutte le conoscenze utili alla depurazione dei dati *panel* da tali fenomeni, focalizzando l'attenzione sui seguenti aspetti:

- correzione della mancata risposta dovuta a *fattori demografici* (decessi, mobilità sul territorio (cambio di domicilio), altri motivi di uscita dal nucleo familiare, ecc.)
- correzione della mancata risposta dovuta a *fattori non demografici* (numero di telefono non più esistente, rifiuto, scarsa sensibilità rispetto agli argomenti oggetto dell'indagine, ecc.)
- necessità di produrre delle stime *panel* coerenti con le popolazioni sezionali di riferimento PLUS (RCFL) 2005 e 2006;
- produzione di stime *panel* per fenomeni di interesse primario per l'indagine (*transizioni tra stati occupazionali e tipologie contrattuali*).

Nel seguito vengono illustrate nel dettaglio le scelte operate nell'ottica del raggiungimento di tali finalità.

4.1. La definizione del file di riferimento e lo studio della mancata risposta panel

[27] Vedi Falorsi (2001).
[28] In contrapposizione a quello di tipo *storico* (o retrospettivo), che fissa la popolazione longitudinale in corrispondenza del limite superiore dell'arco temporale di riferimento.

Come accennato nel precedente paragrafo la scelta operata ai fini dell'individuazione della popolazione longitudinale per il *panel* PLUS è stata quella di considerare come riferimento la popolazione 2005 (popolazione al tempo iniziale). Questo ha comportato la necessita di limitare l'osservazione ai soli individui *compresenti*, cioè coloro che, in target d'indagine nel 2005, potevano con una data probabilità (non nulla) rientrare nella popolazione 2006. A partire dal campione PLUS 2005 si è così costruito un file di riferimento al netto delle *uscite certe* dalla popolazione, costituite dalle persone di 64 anni compiuti nel 2005. Inoltre si è reso necessario un intervento specifico per le casalinghe tra i 40 e i 49 anni, escluse nel 2006 dalla popolazione di riferimento.

Mentre per i primi si è proceduto ad un'eliminazione dal file di studio per l'identificazione del *modello di mancata risposta panel*, nel caso delle casalinghe tra 40 e 49 anni si è tenuto conto del fatto che alcune delle donne intervistate nel 2005 in questa condizione e fascia d'età potessero rientrare nel campione 2006 attraverso una variazione della loro condizione. Di conseguenza, si è pensato di eliminare dal file di riferimento un numero di interviste tali da generare un tasso di risposta omogeneo a quello della classe d'età immediatamente precedente (30-39 anni), valutata essere la "più vicina" – per caratteristiche, necessità e scelte di entrata nel mercato del lavoro – a quella oggetto di studio. Tale condizione è stata imposta all'interno di strati definiti dalla macro area geografica (Nord, Centro e Sud) e dalla tipologia di comune (metropolitano o non metropolitano).

La tabella 6 contiene il risultato dell'allocazione delle casalinghe 2005 tra 40 e 49 anni tra i due collettivi delle non rispondenti 2006 perché *fuori target* o per *altri motivi*. L'effetto ottenuto può essere così sintetizzato: mentre la situazione osservata (o) prevedeva complessivamente un tasso di risposta *panel* del 15% (rispetto al 64,3% della classe d'età delle trentenni) il risultato della procedura di imputazione (i) ha comportato la riallocazione negli strati di quote di donne dalla categoria di *non intervistato per altri motivi* a *non intervistato perché fuori target*, in modo tale che il tasso di mancata risposta per *altri motivi* coincidesse con quello della classe d'età precedente (35,7%).

Tabella 6 – Standardizzazione del campione di riferimento ai fini dello studio della mancata risposta *panel*: ripartizione delle donne casalinghe (40 e 49 anni) tra tipologie di non intervista 2006

	(1) Intervistato	(2) Non intervistato *fuori target*	(3) Non intervistato *altri motivi*	Totale	(1) %	(2) %	(3) %
TOTALE - Italia							
30 a 39 anni	1.484	1	823	2.308	64,3		35,7
40 a 49 anni (o)	378	1	2.142	2.521	15,0		85,0
40 a 49 anni (i)	378	1.243	899	2.521	15,0	49,3	35,7

Fonte: Elaborazioni Isfol-PLUS 2006

(o): osservati – (i): imputati.

È da sottolineare che quanto fatto per il collettivo di donne casalinghe 2005 poteva essere esteso a tutte quelle categorie di individui "a margine di target 2005" che, con maggiore probabilità (a priori), sarebbero usciti dalla popolazione a distanza di un anno[29]. Si è preferito, invece, agire direttamente solo per quelle categorie di intervistati che per motivi di pianificazione di indagine erano certamente (come le persone di 64 anni) o con molte probabilità (come le casalinghe tra 40 e 49 anni) destinate ad uscire. La loro inclusione nel file di riferimento per lo studio della mancata risposta *panel* avrebbe reso il risultato dell'analisi alterato da eventi generati da scelte imposte esogenamente dall'esterno[30]. Sintetizzando, le uscite dal campione *per fuori target* sono state eliminate dal file di riferimento per lo studio della mancata risposta *panel* che è stato così ridotto da 40.386 a 38.170 individui, con una quota *panel* del 64,5%.

4.2. Calibrazione e peso longitudinale finale[31]

La scelta della popolazione longitudinale (riferimento all'anno di partenza del *panel*) ha imposto di considerare come peso base quello sezionale di PLUS 2005. In questo paragrafo è invece presentato nel dettaglio il lavoro fatto al fine di correggere l'autoselezione del campione attraverso un correttore di calibrazione longitudinale in grado di:

- tenere conto del modello sottostante di mancata risposta

[29] Si pensi, ad esempio, ai giovani maschi di 29 anni nella condizione di studente.

[30] Dobbiamo inoltre precisare che la scelta effettuata di adeguare la mancata risposta panel delle casalinghe tra 40 e 49 a quelle della classe d'età immediatamente inferiore non era l'unica possibile e, qualcuno potrebbe affermare, meno robusta di altre. Il punto fondamentale è che, come nel caso delle altre categorie di intervistati 2005 "a margine di target", una parte importante di correzione del *selection bias* è stata affidata alla procedura di calibrazione descritta nel dettaglio nel sottoparagrafo successivo.

[31] Per una trattazione sulle problematiche e soluzioni da adottare nella ponderazione dei dati provenienti da rilevazioni longitudinali vedi Montanari (2001).

- garantire l'affidabilità delle stime prodotte per alcune transizioni notevoli, ovvero di interesse primario per l'indagine.

Questi due obiettivi sono stati raggiunti attraverso l'individuazione di 80 sottopopolazioni, omogenee al loro interno rispetto ad alcune caratteristiche, sulle quali individuare i totali noti da imporre come vincoli di calibrazione. Per l'identificazione degli 80 collettivi ha assunto una notevole importanza l'utilizzo di un innovativo strumento di classificazione non parametrica - *regression tree analysis*. Essi sono il risultato di un numero considerevole di analisi effettuate al fine di individuare le variabili indipendenti maggiormente correlate con la mancata risposta *panel* e le realizzazione di numerosi e significativi eventi specifici quali: transizioni (permanenze) tra (in) condizioni occupazionali, tipologie contrattuali, tipo di lavoro[32].

A. Il primo modello studiato è stato quello della mancata risposta *panel*, con variabile dipendente binaria associata ai due eventi:

$$\begin{cases} 0 = \text{intervistato 2005 non rispondente panel} \\ 1 = \text{intervistato 2005 rispondente panel} \end{cases}$$

e con file di 38.170 individui derivato dal campione complessivo 2005 come dettagliatamente descritto nel precedente sottoparagrafo. Il pacchetto SPSS utilizzato – C&RT – ha permesso di identificare un numero circoscritto di nodi notevoli su cui basare la derivazione del correttore di calibrazione del peso longitudinale.

La tabella 7 evidenzia le percentuali di risposta *panel* per la variabile di indagine *condizione occupazionale* 2005.

Tabella 7 – Incidenza della quota *panel* per condizione PLUS 2005

Condizione 2005	Non intervistato	Panel	Totale	Totale (obs.)
Occupato	35,9	64,1	100,0	16.273
In cerca di lavoro	42,6	57,4	100,0	5.990
Pensionato da lavoro	29,2	70,8	100,0	4.982
Casalinga (altro inattivo)	46,1	53,9	100,0	4.533
Studente	24,5	75,5	100,0	6.392
Total	35,4	64,6	100,0	38.170

Fonte: Elaborazioni Isfol-PLUS 2005-2006

[32] I risultati sono stati sempre supportati da tecniche di regressione logistica di tipo *stepwise (forward) regression*, soprattutto ai fini della selezione delle variabili da inserire come possibili predittori nei modelli non parametrici (*tree analysis*).

È evidente l'eterogeneità della propensione alla risposta osservata: tra le categorie maggiormente rispondenti (con un percentuale superiore alla media del 64,6%) troviamo quella dei pensionati da lavoro e degli studenti, mentre tra i meno rispondenti si collocano i disoccupati e le casalinghe. Anticipando il risultato della procedura è ovvio che il peso finale longitudinale di riporto all'universo debba tenere conto di tale eterogeneità di risposta, attraverso una correzione del peso base tale da attribuire – mediamente – "meno rappresentatività" ad un'informazione di flusso relativa agli studenti e "più importanza", invece, a quella desumibile dalle risposte fornite dagli individui in cerca di lavoro o dalle casalinghe.

Il risultato di questo primo *step* preparatorio della procedura di calibrazione è sintetizzato nella tabella 8. Lo studio ha permesso di individuare tre sole variabili esplicative che, opportunamente classificate ed (endogenamente) incrociate, hanno dato origine a sette collettivi utili alla definizione dei primi totali noti. In ordine di importanza decrescente, sono risultate essere maggiormente correlate con la variabile dipendente : 1) la condizione occupazionale, 2) il titolo di studio e 3) la classe d'età.

Nel dettaglio osserviamo che la percentuale del 64,6% di rispondenti *panel* è scomponibile in percentuali variabili tra il 73,4% di risposta media per studenti e pensionati fino al 36,3% di risposta avuta da persone in cerca di lavoro o casalinghe con titolo di studio medio-basso ed età superiore ai 39 anni.

Tabella 8 - Incidenza di risposta *panel* per gruppi di individui: risultato dell'applicazione di tecniche di regressione non parametrica

Composizione nodi terminali		Numerosità	% di rispondenti *panel*
1	*Studenti e Pensionati da lavoro*	11.374	73,4
2	*Occupati* con titolo di studio pari (o inf.) alla licenza media	3.481	59,8
3	*Occupati* con titolo di studio superiore alla licenza media	12.792	65,2
4	*In cerca di lavoro o casalinghe* con titolo di studio pari (o inf.) alla licenza media ed età inferiore a 40 anni	2.498	57,8
5	*In cerca di lavoro o casalinghe* con titolo di studio pari (o inf.) alla licenza media ed età superiore o pari a 40 anni	1.695	36,3
6	*In cerca di lavoro o casalinghe* con titolo di studio superiore alla licenza media ed età inferiore a 40 anni	4.878	64,9
7	*In cerca di lavoro o casalinghe* con titolo di studio superiore alla licenza media ed età superiore o pari a 40 anni	1.452	44,9
	Totale	38.170	64,6

Fonte: Elaborazioni Isfol-PLUS 2005-2006

Tre sono le caratteristiche positive possedute dai risultati dalla tabella 8:

1. l'identificazione di poche variabili indipendenti in grado di spiegare in modo accurato quella dipendente (variabile)
2. la selezione endogena di tre variabili di base per la stratificazione campionaria (condizione ed età) e calibrazione sezionale (condizione, età e titolo di studio)
3. la presenza della variabile condizione occupazionale, utile ai fini della "quadratura" della procedura, ottenuta attraverso l'implementazione del successivo passo B.

B. Con il secondo gruppo di modelli si è focalizzata l'attenzione sullo studio di alcune "transizioni (o permanenze) notevoli" di importanza primaria per l'indagine. Il fine ultimo è stato quello di:

a) identificare, attraverso *regression tree analysis*, le variabili maggiormente correlate con ciascuno degli eventi di flusso riportati nella tabella 9

b) identificare dei gruppi omogenei – e tra loro il più possibile eterogenei – in base a specifiche caratteristiche, in grado di spiegare quegli eventi nel modo più accurato possibile

c) utilizzare successivamente questi insiemi come vincoli di calibrazione del tipo descritto nel paragrafo 3.2 per il caso dei pesi di riporto all'universo sezionale.

Tabella 9 – Permanenze e transizioni di interesse primario dell'indagine PLUS 2006

Permanenze	**Note**
In cerca di lavoro (disoccupati)	Distintamente per le persone in cerca nel 2005 appartenenti ai gruppi 4&5 e 6&7 della tabella 8
Occupato	Distintamente per i gruppi 2 e 3 di occupati nel 2005 della tabella 8
Occupato a tempo determinato	
Occupato full-time	
Occupato lavoro *tipico*[33]	
Transizioni	
In cerca di lavoro → Occupato	Distintamente per le persone in cerca nel 2005 appartenenti ai gruppi 4&5 o 6&7 della tabella 8
Studente → Attivo (occupato o in cerca)	Solo per gli studenti nel 2005 appartenenti al gruppo 1 della tabella 8
Studente → Occupato	
Casalinga → Attivo (occupato o in cerca)	Distintamente per le casalinghe nel 2005 appartenenti ai gruppi 4&5 o 6&7 della tabella 8

Fonte: Isfol PLUS 2005-2006

[33] Lavoro dipendente con contratto a tempo indeterminato e attività di lavoro autonomo (al netto delle forme contrattuali di parasubordinazione).

Complessivamente (modello A e modelli B) sono state identificate 80 (7+73) sottopopolazioni di riferimento su base 2005, per un totale di 18 regressioni non parametriche. Inoltre sono state garantite le seguenti caratteristiche:

- assenza di collettivi ridondanti (ossia, nessun insieme derivabile come unione di altri)
- collettivi parzialmente sovrapposti tra loro
- contenenti individui con caratteristiche correlate alle variabili di interesse dell'indagine (transizioni e permanenze tra/in condizioni specifiche)
- di numerosità non eccessivamente bassa[34]
- in grado di contribuire alla definizione di un peso *panel* ottenuto con una metodologia coerente con la derivazione dei pesi di riporto all'universo di tipo sezionale.

Il peso finale di riporto all'universo longitudinale è stato quindi derivato attraverso la stessa procedura descritta nel paragrafo 3.2, implementando la formula

$$w_p = w_{0,p} \cdot \gamma_p$$

dove $w_{0,p}$ è in questo caso definito dal peso sezionale (finale) PLUS 2005 – limitatamente alla sola componente campionaria longitudinale – e γ_p è il correttore di calibrazione derivato attraverso l'imposizione degli 80 vincoli di calibrazione sopra descritti sull'intera popolazione di riferimento.

5. Analisi della varianza delle stime

Obiettivo di questo paragrafo è la descrizione della procedura di calcolo dell'attendibilità delle stime utilizzata sull'Indagine PLUS 2006 a seguito delle scelte fatte relativamente al disegno di campionamento e alla costruzione dei pesi di espansione all'universo. Si ritiene infatti indispensabile, per l'utilizzatore dei dati, fornire indicazioni circa l'accuratezza delle stime, in particolare su ciò che concerne l'incertezza delle stesse, dovuta al fatto che nelle indagini campionarie viene rilevata una porzione molto piccola della popolazione di riferimento.

Il software utilizzato per il calcolo degli errori di tipo campionario è Genesees (GENEralised software for Sampling Errors Estimationin Surveys - Versione 3.0), applicativo SAS realizzato e rilasciato gratuitamente dall'Istat (Istituto Nazionale di Statistica). La metodologia implementata nel software Genesees permette di considerare tutti gli stimatori utilizzati nelle indagini campionarie su larga scala come casi particolari degli stimatori di calibrazione (Deville, J. C., Särndal, C. E., 1992). Attraverso l'utilizzo della funzione di Genesees *Analisi dei modelli* è inoltre

[34] Generalmente non inferiore a 50 unità.

possibile costruire dei modelli per la presentazione sintetica degli errori di campionamento.
Rimandando al *Manuale utente ed aspetti metodologici*[35] del software la formulazione teorica della stima della varianza dello stimatore (che nel nostro caso è inserito nell'ampia classe degli stimatori di regressione generalizzata), di seguito verranno illustrate le implicazioni pratiche di tale teoria e i risultati ottenuti sui dati derivanti dalla rilevazione PLUS.
Le informazioni di input necessarie al calcolo dell'attendibilità delle stime sono le seguenti:

1. variabili di interesse (rappresentative) per le quali si desidera misurare la varianza delle stime
2. informazioni relative al disegno di campionamento (stratificazione, pesi base)
3. informazioni relative allo stimatore (pesi finali, variabili ausiliarie, domini pianificati)
4. domini di stima pianificati e non pianificati, rispetto ai quali si desidera conoscere l'errore campionario delle variabili di interesse di cui al punto 1.

I punti 2 e 3 sono stati ampiamente descritti nei paragrafi precedenti, in questo paragrafo ci soffermeremo, invece, sulle scelte effettuate rispetto ai punti 1 e 4.
Per ciò che concerne il primo punto, sono state selezionate 51 variabili di interesse. La scelta di tali variabili (che per brevità non presentiamo) è stata effettuata rispondendo ad una tripla esigenza:

- individuare variabili di interesse fondamentali per la diffusione dei risultati della rilevazione PLUS (es: *tipo di contratto*)
- tenere conto dei diversi moduli che compongono il questionario e che vengono somministrati a diversi profili di rispondenti relativamente a condizione occupazionale, età e sesso
- selezionare delle variabili qualitative le cui distribuzioni di frequenza fossero il più possibile eterogenee tra loro in modo da coprire tutti i livelli assoluti di valori da stimare.

La tabella 10 mostra le partizioni della popolazione di riferimento che definiscono i domini all'interno dei quali sono stati calcolati gli errori campionari delle stime relativamente alle variabili di interesse. Come è possibile notare, i domini di stima sono definiti dalle variabili che concorrono alla stratificazione dell'universo di riferimento nella strategia campionaria. Inoltre, la scelta della variabile *popolazione target* tiene conto dei domini di stima di interesse dell'indagine (precedentemente definiti al paragrafo 2), che si determinano combinando le variabili *sesso*, *età* e *condizione occupazionale* e che in parte si sovrappongono tra di loro. Questa problematica si risolve facilmente considerando gli ultimi tre domini (cioè quelli che vanno a sovrapporsi agli altri) soltanto nella partizione creata con la variabile *condizione occupazionale*.

[35] http://www.istat.it/strumenti/metodi/software/produzione_stime/genesees/index.html#documentazione

Tabella 10 – Domini di stima considerati per il calcolo dell'attendibilità delle stime dell'indagine PLUS 2006

Variabili che definiscono i domini di stima	Modalità
Regione	Tutte le regioni (insieme Valle d'Aosta e Piemonte)
Condizione occupazionale	Occupato/a, In cerca di occupazione, Studente/essa, Casalinga, Pensionato/a
Popolazione target	Occupati 15-29 anni, Studenti 15-29 anni, In cerca di occupazione 15-29 anni, Donne attive 20-39 anni, Donne inattive 20-39 anni, Attivi 50-64 anni, Pensionati*, In cerca di occupazione*, Occupati*

Fonte: Isfol PLUS 2006

* Domini presenti nella variabile "Condizione occupazionale"

Come esempio dei risultati ottenuti, la tabella 11 contiene l'attendibilità della variabile *tipo di contratto* ad un livello di confidenza pari al 95%. Tuttavia, la presentazione del livello di precisione di tutte le stime dell'indagine PLUS sarebbe troppo oneroso e di non facile consultazione per il lettore. Tali difficoltà, comuni a tutte le indagini campionarie complesse, sono state superate grazie all'implementazione di metodi approssimati (Manuale utente ed aspetti metodologici di Genesees 3.0) che consentono la definizione di *modelli regressivi* che mettono in relazione ciascuna stima con il proprio errore di campionamento.

Tali modelli vengono formalizzati con la seguente espressione:

$$\widetilde{CV}(\widetilde{Y}_{REG}) = \sqrt{\exp\left[\widetilde{b}_0 + \widetilde{b}_1 \log\left(\widetilde{Y}_{REG}\right)\right]}$$

dove $\widetilde{Y}_{REG}$ è la stima della variabile oggetto d'interesse, mentre $\widetilde{b}_0$ e $\widetilde{b}_1$ sono i parametri del modello[36]. Applicando tale soluzione, nella tabella 12 sono riportati i valori dei parametri del modello di interpolazione ed il relativo coefficiente di determinazione (R^2) che permettono di calcolare, in modo autonomo, i Coefficienti di Variazione ($\widetilde{CV}$) di ciascuna stima per ogni dominio di stima pianificato.

Nel tabella 13 sono invece illustrati i livelli delle stime, per ciascun dominio, corrispondenti a determinati valori del $\widetilde{CV}$ ad un livello di confidenza pari al 95%. E' evidente che l'errore campionario di una singola stima può essere calcolato all'interno di più domini pianificati: ad esempio il Coefficiente di Variazione ($\widetilde{CV}$) della stima del numero di occupati nel Lazio può essere calcolato applicando il modello di interpolazione relativo alla regione Lazio e quello relativo agli

[36] Per un maggiore dettaglio teorico, vedi il paragrafo 9.4 del Rapporto PLUS 2005.

occupati ottenendo risultati diversi. Si consiglia pertanto di utilizzare con prudenza lo strumento appena presentato considerando, ogni qual volta non è possibile scegliere il dominio di stima più idoneo e stringente, il livello di errore più alto in modo da evitare interpretazioni approssimative.

Tabella 11 - Stima e corrispondente errore campionario percentuale per tipo di contratto (livello di confidenza: 95%)

Tipo di contratto	Totale	
	Stima	Errore %
Non occupato	11.385.273	0,0
Lavoro a tempo indeterminato	14.253.628	0,6
Lavoro a tempo determinato (escluso CFL, apprendistato, inserimento)	1.075.122	5,4
Contratto formazione lavoro (CFL)	133.822	12,2
Apprendistato	346.912	6,1
Contratto d'inserimento	180.425	11,1
Lavoro interinale o a somministrazione	147.575	13,1
Job sharing o lavoro ripartito	8.876	47,5
Lavoro intermittente o a chiamata	157.950	17,6
Collaborazioni coordinate e continuative (Co.Co.Co.)	375.176	8,9
Collaborazione occasionale (Ritenuta d'acconto)	358.661	14,8
Lavoro a progetto	559.561	7,0
Titolare d attività - Imprenditore	2.429.413	3,4
Associati in partecipazione	63.810	24,8
Attività in proprio (Partita IVA)	1.641.244	4,2
Coadiuvante familiare	147.215	13,8
Stage, Alternanza scuola - lavoro	33.988	24,0
Pratica professionale	55.970	17,4
Tirocinio	33.666	17,1
Altro Dipendente	423.798	9,2
Altro Autonomo	191.699	15,8

Fonte: Elaborazioni Isfol-PLUS 2006

Tabella 12 -Valori dei coefficienti B0, B1 e dell'indice di determinazione R2 (%) dei modelli utilizzati per il calcolo degli errori campionari delle stimedi frequenze assolute per dominio di stima

DOMINIO	Parametri del modello		
	B0	B1	R2
Nazionale	8,388507	-1,09178	87,16341
Regione			
Valle d'Aosta, Piemonte	7,852074	-1,09956	85,04383
Liguria	7,14382	-1,08884	77,99063
Lombardia	8,021716	-1,0634	83,61206
Trentino Alto Adige	6,272252	-1,00989	75,09371
Veneto	8,710655	-1,1468	82,82349
Friuli Venezia Giulia	6,3725	-1,02833	80,57017
Emilia Romagna	8,011561	-1,11688	82,48477
Toscana	8,110051	-1,10716	78,01439
Marche	6,888725	-1,05512	81,92938
Umbria	5,522101	-0,91245	61,30315
Lazio	8,715298	-1,13548	88,68529
Molise	5,740516	-1,06228	84,53297
Abruzzi	7,980218	-1,19048	80,99534
Campania	10,40718	-1,33546	84,69706
Puglia	8,329	-1,13198	85,41354
Basilicata	6,388192	-1,05984	79,58452
Calabria	8,166181	-1,14014	83,16082
Sicilia	8,660947	-1,15283	89,0594
Sardegna	6,969155	-1,04451	81,92703
Condizione occupazionale			
Occupato/a	9,994645	-1,17333	89,08139
In cerca di lavoro	8,128095	-1,13258	87,06846
Pensionato/a da lavoro	7,949919	-1,12838	93,161
Casalinga	9,530963	-1,28229	86,63646
Studente/essa	9,013568	-1,25517	89,38156
Popolazione target			
Occupati 15-29 anni	9,476983	-1,23902	92,39538
Studenti 15-29 anni	10,95277	-1,43153	85,05505
In cerca 15-29 anni	8,903035	-1,20716	94,86613
Donne attive 20-39	8,480286	-1,12701	83,26214
Donne inattive 20-39	8,564138	-1,21519	91,29316
Attivi 50-64 anni	8,850168	-1,17058	85,37722

Fonte: Elaborazioni Isfol-PLUS 2006

Tabella 13 - Livelli delle stime per determinati valori del Coefficiente di Variazione (CV) per dominio di stima

DOMINIO	Livello di significatività (valori del CV)			
	0,1	0,15	0,2	0,25
Nazionale	147.473	70.168	41.425	27.526
Regione				
Valle d'Aosta, Piemonte	83.221	39.805	23.588	15.719
Liguria	48.550	23.054	13.591	9.021
Lombardia	143.505	66.939	38.967	25.611
Trentino Alto Adige	47.616	21.331	12.067	7.757
Veneto	110.341	54.405	32.941	22.322
Friuli Venezia Giulia	43.271	19.666	11.239	7.282
Emilia Romagna	80.529	38.961	23.275	15.609
Toscana	97.198	46.726	27.789	18.570
Marche	53.822	24.956	14.466	9.477
Umbria	66.107	27.181	14.468	8.872
Lazio	124.381	60.896	36.688	24.765
Molise	16.969	7.909	4.601	3.023
Abruzzi	39.015	19.742	12.176	8.369
Campania	76.213	41.526	26.990	19.323
Puglia	91.686	44.790	26.943	18.164
Basilicata	31.974	14.876	8.644	5.673
Calabria	73.246	35.966	21.713	14.680
Sicilia	99.456	49.220	29.880	20.289
Sardegna	64.930	29.873	17.220	11.233
Condizione occupazionale				
Occupato/a	253.495	127.002	77.776	53.169
In cerca di lavoro	76.325	37.300	22.443	15.134
Pensionato/a da lavoro	67.957	33.122	19.892	13.394
Casalinga	61.334	32.588	20.806	14.690
Studente/essa	51.539	27.012	17.079	11.969
Popolazione target				
Occupati 15-29 anni	86.298	44.849	28.189	19.663
Studenti 15-29 anni	52.471	29.778	19.923	14.587
In cerca 15-29 anni	72.411	36.988	22.965	15.867
Donne attive 20-39	110.274	53.701	32.230	21.691
Donne inattive 20-39	50.881	26.106	16.260	11.262
Attivi 50-64 anni	98.188	49.113	30.042	20.519

Fonte: Elaborazioni Isfol-PLUS 2006

5.1. Una sperimentazione per lo sviluppo di una procedura generalizzata di analisi della varianza delle stime

I dati dell'indagine PLUS trattano principalmente variabili qualitative, per cui i modelli stimati fanno riferimento essenzialmente a stime di frequenze assolute.

La metodologia resta comunque valida in caso si voglia stimare una frequenza relativa o un qualsiasi indicatore riferiti all'intera popolazione di riferimento del dominio o ad un altro totale noto tra

quelli considerati nella fase di poststratificazione del campione. In tal caso il denominatore, il totale della popolazione, non viene considerato affetto da errore in quanto costituisce un valore noto. Nel caso si voglia calcolare l'errore relativo in una sottopopolazione diversa, ad esempio la popolazione che presenta una certa modalità di una variabile di interesse, è necessario ricorrere ad una approssimazione. Infatti, la stima di una frequenza relativa o di un qualunque indicatore riferita ad un sottogruppo di persone, è ottenibile come rapporto tra due quantità entrambe stimate:

$$\hat{R}_d = \frac{\hat{N}_d}{\hat{D}_d}$$

Una valutazione approssimata dell'errore relativo della stima $\hat{R}_d$, sotto l'ipotesi di incorrelazione tra $\hat{N}_d$ e $\hat{D}_d$ si può ottenere come:

$$\hat{\varepsilon}(_d\hat{R}) = \sqrt{\hat{\varepsilon}^2(\hat{N}_d) - \hat{\varepsilon}^2(\hat{D}_d)}$$

Per consentire una valutazione sistematica dell'attendibilità delle stime ottenute dall'indagine attraverso la metodologia finora descritta, necessaria soprattutto nella fase di validazione dei risultati, è stata realizzata una procedura generalizzata in grado di fornire indicazioni sulla precisione di una qualsiasi stima. La procedura implementata si basa sulla generalizzazione della metodologia fin qui descritta, e consente il calcolo degli errori campionari per le stime ottenute con autonome elaborazioni dei dati elementari dell'indagine.

La procedura si articola in distinte funzionalità e fa riferimento ai tre casi in cui si vanno a costruire gli intervalli di confidenza:

1. stime di frequenze assolute
2. stime di rapporti in cui il numeratore è una stima ed il denominatore un totale noto
3. stime di un rapporto in cui numeratore e denominatore sono entrambi stimati.

La scelta del metodo di stima corretto dell'errore relativo della stima non appare sempre univoca, in particolare rispetto al secondo caso: pur considerando in generale il denominatore costituito da un totale noto, quindi non affetto da errore campionario, si potrebbe però essere interessati ad una diversa classificazione o all'analisi di tale indicatore rispetto ad una specifica sottopopolazione. In questo caso la procedura è in grado di valutare se il filtro applicato o la classificazione richiesta non facciano più coincidere il denominatore con il relativo totale noto ed applicare quindi la necessaria approssimazione dell'errore valida per gli stimatori rapporto.

La base informativa su cui si poggia la procedura infatti è costituita da un set di metadati operativi che comprendono una generalizzazione della formula di calcolo degli indicatori, una loro classificazione che consente di applicare il metodo di calcolo corretto della stima dell'errore relativo e tutti i vincoli definiti nella procedura di poststratificazione, oltre a tutte le informazioni normalmente necessarie per l'elaborazione dei dati elementari: tipologia delle variabili, classificazioni ed etichette.

L'applicazione è stata sviluppata in SAS Macro Language utilizzando anche componenti web ed è attualmente disponibile all'interno della rete intranet per l'utenza Isfol; è in corso l'estensione ad altre indagini campionarie sia svolte dall'istituto che da altri istituti di ricerca, nonché la definizione delle modalità per la sua diffusione agli utenti esterni che richiedono l'acquisizione dei dati elementari dell'indagine.

6. Correzione ed imputazione delle mancate risposte parziali

6.1. Imputazione delle mancate risposte per i redditi da lavoro

Tra le informazioni disponibili annualmente nell'indagine PLUS, un ruolo centrale è occupato dai redditi da lavoro. La domanda prevista nel questionario 2006 non ha subito variazioni rispetto alla precedente *wave*[37]. Sono state poste domande distinte agli occupati con contratto di lavoro alle dipendenze, ai lavoratori autonomi e ai collaboratori[38] (ai dipendenti è stato chiesto il reddito netto mensile, agli autonomi il reddito lordo annuo ed ai collaboratori il reddito lordo mensile). Per ciascuna delle tre macro categorie di lavoratori è stata operata una correzione dei valori anomali (*outliers*) che non prevede l'eliminazione di alcuna osservazione, limitando al minimo la perdita di informazioni disponibile nel dataset. In generale, tutti i redditi dichiarati al di sopra e al di sotto di *due specifiche soglie* sono stati corretti e posti uguali a tali valori[39]. Ciascuna delle tre categorie di reddito ha richiesto un opportuno criterio per l'identificazione delle soglie limite da applicare: per i lavoratori dipendenti e i collaboratori sono state considerate due soglie, una inferiore e una superiore, fissate rispettivamente ad un ventesimo e a dieci volte il valore della mediana della distribuzione dei redditi pertinente; per i redditi dichiarati dai lavoratori autonomi sono state scelte delle soglie diverse dalla precedenti, a causa della loro particolare forma distributiva (caratterizzata da un elevato grado di dispersione) ed in funzione dalla tipologia di lavoro (attività

[37] In PLUS 2008 la qualità dell'informazione sui redditi degli individui dovrebbe essere migliore di quella avuta nelle due precedenti indagini, per l'introduzione di alcune domande integrative (auto-collocazione dei soggetti in classi di reddito nel caso di rifiuto nella dichiarazione dell'ammontare esatto, indicazione del reddito complessivo familiare, capacità di spesa familiare distinta per beni di prima necessità ed eventi imprevisti, ecc.) in grado di permettere attribuzioni *ex-post* molto più precise e supportate da un'ampia informazione aggiuntiva.

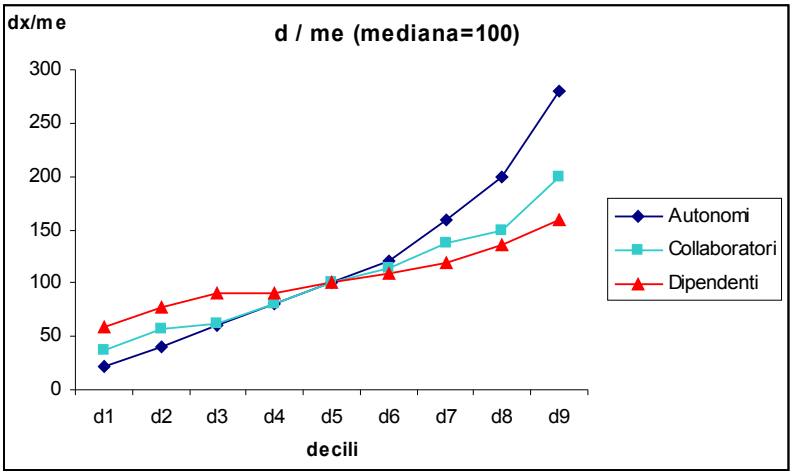
[38] Categoria comprensiva dei co.co.co., lavoro a progetto e collaborazioni occasionali.

[39] Questa procedura è utilizzata, ad esempio, in abito LIS (*Luxembourg Income Study*) che costituisce una delle maggiori banche dati internazionali sui redditi.

indipendente) la quale implica delle remunerazioni influenzabili da entrate o perdite di esercizio anche "straordinarie". Per questi le soglie inferiore e superiore sono state fissate rispettivamente pari a un cinquantesimo e 15 volte il valore della mediana.
Il grafico 2 sintetizza "la forma" delle tre diverse distribuzioni in oggetto, in termini di dispersione dei redditi attorno al valore della mediana. I rapporti inter-decili evidenziano, per i lavoratori autonomi, una spezzata sempre inferiore (superiore) alle altre due distribuzioni al di sotto (di sopra) del valore mediano, mentre risultano nettamente meno diseguali le distribuzioni dei collaboratori e dei dipendenti.

La disponibilità di interviste *panel* nel 2006 ha permesso di perfezionare la procedura di imputazione delle mancate risposte parziali per la variabile reddito rispetto a quanto fatto in PLUS 2005, ricorrendo ad un avanzamento nella tecnica standard di imputazione per donatore con schemi di stratificazione iniziale già utilizzata in passato (procedura *hot-deck*). Complessivamente sono stati individuati 9.477 soggetti occupati in entrambi gli anni di rilevazione, di cui 9.264 contemporaneamente in target di domanda[40]. Osservando le percentuali di rispondenti dipendenti, autonomi e collaboratori è stata riscontrata una maggiore propensione alla risposta rispetto alla componente *non panel* oltre ad un cospicuo recupero di non rispondenti 2005 e rispondenti 2006: il 42,6% delle unità *panel* non rispondenti nel 2005 (1.082 su 2.542) hanno infatti dichiarato il loro reddito l'anno successivo.

Grafico 2 – Rapporti inter-decilici, distribuzioni dei redditi da lavoro dipendente, autonomo e per collaboratori, mediana (d5)=100



Fonte: Elaborazioni Isfol-PLUS 2006

[40] Nel 2005 si era preferito non richiedere informazioni sul reddito agli occupati con tipologie contrattuali quali stage, inserimento lavorativo, ecc., oltre che a coloro che dichiaravano di avere un contratto informale o di non ricordare/non sapere.

Se da un lato tale dato assume un rilevante interesse in un'ottica di produzione di stime *panel* quanto più consistenti, dall'altro - ai fini della procedura di imputazione delle mancate risposte parziali - è importante focalizzare l'attenzione sul collettivo di occupati con reddito dichiarato nel 2005 e non dichiarato nel 2006. Limitatamente alla categoria dei dipendenti *panel* 2005-2006[41], la stratificazione del campione necessaria alla definizione delle celle dalle quali estrarre gli *individui donatori* è stata definita anche sulla base del quartile di reddito di appartenenza dell'individuo nel 2005.

Ricordiamo che la stratificazione del campione, attraverso variabili che si ipotizza siano buoni predittori del processo generatore del dato mancante, consente di affinare la scelta del donatore in modo efficiente, controllando la selezione per gruppi omogenei di unità[42]. La scelta della variabili esplicative – e l'aggregazione delle rispettive modalità – è avvenuta in modo indipendente per le tre tipologie di reddito rilevate dall'indagine (garantendo, comunque, l'utilizzo di variabili demografiche fondamentali quali il sesso, l'età e la ripartizione geografica). Il fine è stato quello di tenere conto delle peculiarità di ciascuna distribuzione (numerosità campionaria, tasso di mancata risposta parziale, possibilità di utilizzare informazione aggiuntiva di tipo *panel*) e, contemporaneamente, di garantire che la condizione di numerosità minima dei donatori nelle celle di stratificazione (fissata a 5 unità) fosse sempre verificata.
Nella tabella 14 sono elencate le variabili utilizzate nelle tre procedure di imputazione, mentre il risultato finale è sinteticamente illustrato nella tabella 15: questa contiene alcune statistiche descrittive di base relative alle distribuzioni di partenza (osservate) e finali (osservate & imputate) relative a 5 categorie di occupati 2006[43].

[41] Non è stato possibile estendere la metodologia panel a tutte le altre transizioni tra contratti (dipendente-autonomo, dipendente-collaboratore, collaboratore-dipendente, ecc.) a causa della bassa numerosità campionaria osservata. Dovendo implementare stratificazioni per importanti (e irrinunciabili) variabili di controllo, questo avrebbe compromesso l'applicabilità della procedura *hot-deck* descritta di seguito.

[42] Per le ipotesi fondamentali di *missing at random* e *missing completely at random* e per i problemi dovuti al "vincolo di pignorabilità" si veda Little and Rubin (1987) e il Rapporto PLUS 2005.

[43] La tabella contiene statistiche descrittive ottenute con dati non pesati, in quanto l'applicazione del peso alla distribuzione osservata non avrebbe senso in questo contesto, visto che la popolazione di occupati risulterebbe incompleta.

Tabella 14 - Variabili di stratificazione utilizzate nelle tre procedure di imputazione

Variabile da imputare	Variabili indipendenti di stratificazione (e modalità)
Reddito da lavoro dipendente	• sesso (maschi, femmine) • età in classi 4 (15-29, 30-39, 40-49, 50-64) (15-39, 40-64) per le donne part-time • area geografica (nord, centro, sud e isole) • istruzione (fino a licenza media, diploma, laurea e oltre) • tipo lavoro 1 (part-time, full-time) • [*solo quota panel*] quartili della distribuzione del reddito dei dipendenti 2005 (Q1, Q2, Q3, Q4)
Reddito da lavoro autonomo	• sesso (maschi, femmine) • età in classi 2 (fino a 29, 30-65) • area geografica (nord, centro, sud e isole) • settore (agricoltura, industria & costruzioni, commercio, servizi)
Reddito per collaboratori	• sesso (maschi, femmine) • età in classi 3 (15-29, 30-49, 50-64) • area geografica (nord, centro, sud e isole) • istruzione (fino a licenza media, diploma, laurea e oltre)

Fonte: Isfol PLUS 2006

Tabella 15 - Statistiche descrittive per le distribuzioni originarie e finali (con mancate risposte imputate) per tipologia di lavoratore

	Mean	Median	CV	Skewness	Kurtosis	P75/P25
Autonomi						
Reddito lordo percepito nell'anno 2005	37.338	25.000	1,5	7,0	72,2	3,1
Reddito lordo percepito nell'anno 2005 - Imputato	31.402	20.000	1,1	3,6	20,8	3,6
Collaboratori						
Reddito lordo percepito nell'ultimo mese	1.034	800	2,1	22,1	584,8	2,4
Reddito lordo percepito nell'ultimo mese - Imputato	1.024	800	0,9	4,2	30,3	2,4
Dipendenti						
Reddito netto percepito nell'ultimo mese	1.221	1.100	0,6	23,7	1.272,8	1,6
Reddito netto percepito nell'ultimo mese - Imputato	1.212	1.100	0,5	3,8	33,3	1,6
Dipendenti full-time						
Reddito netto percepito nell'ultimo mese	1.323	1.200	0,6	25,6	1.334,7	1,5
Reddito netto percepito nell'ultimo mese - Imputato	1.320	1.200	0,5	4,4	40,1	1,5
Dipendenti part-time						
Reddito netto percepito nell'ultimo mese	705	650	0,4	1,3	5,8	1,7
Reddito netto percepito nell'ultimo mese - Imputato	726	700	0,4	1,9	12,0	1,7

Fonte: Elaborazioni Isfol-PLUS 2006

Tra le caratteristiche fondamentali osserviamo che:

- la media e la mediana diminuiscono considerevolmente per il collettivo degli autonomi, rimangono pressoché invariate per le distribuzioni dei dipendenti full-time e dei collaboratori, diminuiscono entrambe lievemente per i dipendenti part-time
- il coefficiente di variazione (CV) diminuisce *sempre* a seguito dell'imputazione
- vengono ridotte considerevolmente l'asimmetria e la curtosi delle distribuzioni osservate
- il rapporto tra i percentili 75° e 25° rimane pressoché invariato – tranne che per una leggera variazione nella distribuzione degli autonomi – indicando un'azione "riequilibrante" dell'imputazione rivolta soprattutto all'aggiustamento delle code estreme delle distribuzioni.

6.2. Imputazione delle mancate risposte per le variabili categoriali

Al momento della consegna da parte della società Doxa, i dati raccolti attraverso la rilevazione presentavano alcune variabili con problemi di mancata risposta parziale. Tale fenomeno, comune alla maggior parte delle indagini campionarie con questionario complesso, può essere causato da molteplici fattori. In particolare, per l'Indagine PLUS 2006, l'origine delle mancate risposte parziali è dovuta:

- al questionario (ad esempio, percorsi dell'intervista molto articolati)
- al rispondente (incapacità o rifiuto di rispondere a specifiche domande)
- al rilevatore (domanda non posta, posta non correttamente, …)
- al processo di rilevazione (errore di codifica o di registrazione).

Allo scopo di ottenere un *data-set* senza valori mancanti[44] per l'applicazione di metodi standard di stima e inferenza attraverso l'utilizzo dei pesi di espansione all'universo, sulla quale poggia la strategia campionaria dell'indagine in questione, si è scelto di implementare una procedura di imputazione delle mancate risposte parziali di tipo non parametrica (o *data based*). Tale processo consiste nella sostituzione dei valori mancanti con valori opportunamente determinati sulla base dei valori osservati.

Nel dettaglio, si è utilizzato il metodo dell'imputazione dal donatore più vicino con classi di imputazione determinate attraverso variabili ausiliarie correlate al meccanismo di mancata risposta. Con questo metodo il donatore è stato scelto in modo tale da minimizzare la funzione di distanza multivariata (Indice di Similarità di Gower) all'interno di ogni classe di imputazione, calcolata tra l'unità del campione con mancata risposta e tutte le altre unità senza dati mancanti sulla base di variabili ausiliarie correlate alla variabile da imputare. In caso di due o più osservazioni

[44] Solo in pochi casi è stato deciso di non intervenire con l'imputazione delle mancate risposte parziali, riguardanti soprattutto quei fenomeni (variabili) per i quali non è stato possibile individuare appropriate variabili ausiliarie.

che presentano lo stesso valore minimo di distanza rispetto ad uno stesso record con valore mancante, l'unità donatrice "più vicina" viene selezionata casualmente assegnando ad ognuna di esse la stessa probabilità di scelta. Si è a questo punto utilizzato il valore osservato sull'unità "più vicina" per effettuare l'imputazione.

Tra i vantaggi connessi con questo metodo di imputazione delle mancate risposte parziali va sicuramente citato quello relativo al mantenimento ottimale delle distribuzioni multivariate originali anche grazie alla possibilità, nel caso di indagini su larga scala come PLUS, di trovare uno stesso donatore per predire simultaneamente molte mancate risposte. Uno dei possibili svantaggi può invece essere rappresentato dal fatto che uno stesso donatore può essere utilizzato troppe volte provocando distorsioni di varia entità nella distribuzione delle variabili e sottostima della variabilità delle stime. Tale rischio è stato però controllato nell'implementazione del metodo, i cui risultati non hanno evidenziato casi di donatori sovrautilizzati.

7. Conclusioni

Questo contributo è il risultato del lavoro dei ricercatori Isfol che da qualche anno si occupano della progettazione, coordinamento, implementazione ed analisi dell'indagine PLUS. L'intero processo di produzione di informazione statistica è caratterizzabile dai seguenti aspetti: a) preparazione di un questionario (anche in collaborazione con esperti esterni) che permetta di rilevare fenomeni rari, poco esplorati, ma di notevole interesse nelle analisi di un mercato del lavoro in continua trasformazione; b) organizzazione della fase di rilevazione attraverso interviste CATI realizzate da società esterne, attraverso la partecipazione attiva nella fasi strategiche di formazione degli intervistatori, monitoraggio delle interviste e continui rapporti con i responsabili scientifici e operativi delle società; c) utilizzo di metodologie teoricamente fondate, replicabili nel tempo e, per alcuni aspetti, innovative; d) rigoroso trattamento e omogeneizzazione dei *database*, con il duplice obiettivo di minimizzare la perdita di informazione acquisita e integrare quella assente attraverso l'utilizzo di adeguate procedure di imputazione; e) diffusione ad utenti esterni (previa sottoscrizione di un opportuno protocollo) dei microdati PLUS e di tutti i supporti utili alle loro analisi (tracciati record, questionario, strumento semplificato per la valutazione della significatività delle stime).

Il modello seguito dal gruppo di lavoro PLUS è del tutto equiparabile a quello adottato in ISFOL e in altri istituti di ricerca per alcune indagini campionarie di interesse nazionale. La sua applicabilità a simili progetti di competenza ISFOL è pressoché totale ma, in ogni caso, subordinata alla progettazione di un processo di lavoro che preveda nelle fasi di implementazione, controllo, analisi e diffusione dei dati il rispetto di appropriati standard tecnico-scientifici.

Bibliografia

Centra M., Falorsi P.D. (a cura di), *Strategie di campionamento per il monitoraggio e la valutazione delle politiche*, Roma, Isfol, 2008 (Temi e Strumenti)

Deville J.C., Särndal C.E., *Calibration Estimators in Survey Sampling*, "Journal of the American Statistical Association", vol.87, 1992, pp.367-382

Dorfman A.H., Royall R.M., Valliant R., *Finite Population Sampling and Inference: a Prediction Approach*, New York , John Wiley & Sons, 2000

Falorsi P.D., Falorsi S., Russo A., Tranquilli G.B., *Indagini ripetute nel tempo: obiettivi e disegni di rilevazione*, "Rivista di Statistica Ufficiale", n.1, 2001, pp. 5-11

Horvitz D.G., Thompson D.J., *A generalization of sampling without replacement from a finite universe*, "Journal of the American Statistical Association", n.47, 1952, pp. 663-685

Little R.J.A., Rubin D.B., *Statistical analysis with missing data*, New York, Wiley, 1987

Mandrone E., Radicchia D. (a cura di), *PLUS - Participation Labour Unemployment Survey*, Roma, Isfol, 2006 (I libri del Fondo sociale europeo)

Montanari G.E., *La ponderazione dei dati nelle rilevazioni longitudinali mediante campione panel*, "Rivista di Statistica Ufficiale", n.1, 2001, pp. 27-41

Già pubblicati nella collana Studi Isfol:

Mandrone E., *La riclassificazione del lavoro tra occupazione standard e atipica: l'Indagine Isfol Plus 2006*, Studi Isfol 2008/1

Indiretto G., De Santis A., Addobbo T., Belmonte S., *Fiscalità e offerta di lavoro: una prospettiva di genere*, Studi Isfol 2008/2

Baronio G., Marocco M., *Il Caso dei "Centri integrati per l'impiego": le prospettive di costruzione di un sistema integrato di politiche attive e passive in Italia*, Studi Isfol 2008/3

Fabrini L., Raciti P., Ranieri C., *Un modello di Osservatorio per il governo del sistema delle professioni sociali e lo sviluppo dei servizi alla persona*, Studi Isfol 2008/4

Landi R., *Le procedure di accertamento dello stato di disoccupazione e di attivazione dei disoccupati nei Centri per l'impiego*, Studi Isfol 2008/5

Mandrone E., *Quando la flessibilità diviene precarietà: una stima sezionale e longitudinale*, Studi Isfol 2008/6

Grimaldi A., Barruffi, A., Nucera U., Colombo L., *Le rappresentazioni sociali dell'orientamento: risultati di uno studio pilota*, Studi Isfol 2009/1

Centra M., Cutillo A., *Differenziale salariale di genere e lavori tipicamente femminili*, Studi Isfol 2009/2