capitolo 9

IMPIANTO DELL'INDAGINE PLUS

9.1 L'INDAGINE*

L'indagine campionaria nazionale Isfol Plus (Partecipation, Labour, Unemployment, Survey) è stata ideata dall'area "Ricerche sui sistemi del lavoro" con l'intento di analizzare la composizione di alcuni aggregati, o target, del mondo del lavoro particolarmente interessanti sotto il profilo della loro evoluzione, e quindi degli impatti che creeranno negli equilibri del sistema. Il progetto rientra nelle attività finanziate dal Fondo sociale europeo e gestite in collaborazione con la Direzione generale del mercato del lavoro, l'orientamento e la formazione del Ministero del lavoro e della previdenza sociale.
L'indagine, presente nel Piano statistico nazionale, è sintesi di due grandi filoni di ricerca: quello sociologico e quello economico. In questa rilevazione, infatti, si alternano temi economici, come il reddito e l'occupazione, con quesiti sulla condizione familiare e il contesto in cui vive l'individuo intervistato. La ricchezza della fonte dati risiede principalmente nell'integrazione di ambiti spesso analizzati in maniera disgiunta. Qui l'intento di contestualizzare i problemi del mondo del lavoro con la realtà personale e familiare consente di comprendere legami e relazioni invisibili ad indagini monotematiche.
L'universo di riferimento è la popolazione in età tra i 15 e i 64 anni, pari a 34.779.159 individui; la rilevazione è telefonica di tipo Cati su un campione finale di 40.386 interviste. Il disegno dell'indagine prevede cinque target di riferimento: i giovani tra i 15 e i 29 anni, le donne tra 20 e 49 anni, la popolazione in età compresa tra 50 e 64 anni, le persone non occupate e in cerca di lavoro, le persone occupate.
Per ciascuno dei target sono state definite una serie d'informazioni che l'indagine ha voluto rilevare, necessarie al conseguimento degli obiettivi del progetto, tradotte poi in moduli del questionario:

* E. Mandrone, Isfol, Area "Ricerche sui sistemi del lavoro".

Giovani. Le criticità relative al mercato del lavoro nelle classi giovanili sono numerose: l'indagine ha dato conto dei profili di transizione scuola-lavoro, dei rischi di segmentazione dell'offerta e delle determinanti del fenomeno dell'abbandono scolastico e dell'*overeducation*. L'intento è stato quello di studiare la composizione dei percorsi formativi (istruzione e formazione) e occupazionali, alla luce sia del contesto che della famiglia.

Donne. Alla componente femminile della popolazione, sia attiva che inattiva, è dedicato un ampio modulo del questionario. La formazione dell'offerta di lavoro femminile è determinata sia da fattori di natura economica che da aspetti socioculturali. Sono analizzate perciò sia le caratteristiche legate alla struttura familiare, al reddito di riserva e ai redditi di sostituzione, sia aspetti relativi ai modelli extraeconomici familiari e territoriali. Particolare attenzione è rivolta ai pattern di transizione nell'inattività e di rientro nell'occupazione in occasione dei periodi di maternità. L'indagine rileva infine una misura, seppure solo percepita, dell'offerta di servizi alle famiglie, in particolare quelli di *child-care* e d'assistenza agli anziani. L'intento è quello di porre in luce le scelte di partecipazione dovute a pressioni o condizioni familiari, nel senso di capire come la famiglia, l'humus culturale, i servizi sociali e il contesto economico abbiano posto le condizioni favorevoli o, invece, vincoli non superabili alla partecipazione.

Over 50. La componente della popolazione in età superiore a 50 anni suscita interesse non soltanto per le sue caratteristiche peculiari ma anche nell'ottica del prolungamento della vita attiva. L'indagine rileva le determinanti delle scelte previdenziali e i modelli di uscita dal mercato. Il modulo messo a punto per la popolazione *over* 50 contiene quesiti sulle motivazioni che spingono o hanno spinto gli intervistati all'uscita dal mercato o alla permanenza, con particolare attenzione agli aspetti sia finanziari sia familiari ma anche alle pressioni e alle tensioni provenienti dal posto di lavoro.

Disoccupati/inoccupati. Per quanto attiene alla popolazione in cerca di prima o di nuova occupazione, l'indagine approfondisce la conoscenza relativa all'intensità e alle modalità di ricerca di un lavoro: sono rilevati i canali utilizzati nella ricerca di un impiego e quali hanno permesso di trovare un'occupazione, guardando così sia all'efficacia dei diversi tipi di *match* tra domanda e offerta, sia alla qualità del *match* in termini di caratteristiche dell'occupazione. Una serie di quesiti sono infine dedicati al rapporto con i Centri per l'impiego e alla percezione dei cittadini circa i servizi ricevuti.

Occupati. La sezione relativa agli occupati, accanto agli usuali parametri descrittivi dell'occupazione (settore d'attività economica, qualifica, professione) è stata affiancata da una rilevazione dettagliata delle forme contrattuali, analizzata sia rispetto alle modalità di prestazione del lavoro, sia agli aspetti di sicurezza e di prospettive dell'occupazione. L'indagine rileva inoltre gli elementi che tradizionalmente caratterizzano un rapporto di subordinazione rispetto ad uno autonomo (orari, presenza presso la sede di lavoro, periodicità delle retribuzioni, diritti e tutele sul lavoro); le caratteristiche peculiari delle forme di lavoro non tradiziona-

li (part-time, lavoro interinale, lavoro parasubordinato, contratti a termine); la qualità percepita e le caratteristiche più oggettive dell'occupazione svolta. Tra le informazioni di particolare importanza è stato inserito il reddito da lavoro e la ricostruzione dell'anzianità di servizio e lavorativa.
Particolare attenzione è stata data alla ricostruzione del network familiare dell'individuo, allo scopo di individuare i legami di tipo extraeconomico e le scelte indotte non tanto dal mercato quanto dalla struttura e dai rapporti familiari. Un maggiore livello d'informazione che l'indagine rileva è relativo alle scelte individuali in relazione alle priorità o caratteristiche della coppia, considerando le scelte lavorative una complessa sintesi delle istanze presenti in famiglia.
Il livello formativo e il profilo occupazionale della famiglia d'origine rappresentano informazioni cruciali per analizzare numerosi aspetti del mercato del lavoro e delle scelte formative, in virtù della persistenza intergenerazionale dei livelli d'istruzione acquisiti e delle scelte lavorative. A tal fine l'indagine prevede una dettagliata analisi della struttura della famiglia di provenienza degli intervistati. La motivazione principale delle scelte individuali è la famiglia; è in essa allora che si devono individuare molte delle esigenze che portano alle scelte lavorative; sono queste che spesso determinano l'intensità dell'impegno lavorativo e su di esse ci si deve concentrare per vedere come si collocano gli individui nel mercato. È necessario pertanto capire quando le scelte dell'individuo sono oggetto di mediazione tra esigenze in seno alla famiglia, quando rispondono ad una contrattazione di mercato e quando invece la formazione, le aspirazioni e le abilità soggettive determinano una divergenza dai profili occupazionali attesi.
Si è ritenuto, inoltre, utile identificare la presenza d'individui con ridotta autonomia fisica all'interno del nucleo familiare, per osservare come le scelte individuali sono sostenute o ostacolate dal più ampio contesto socio-economico.
Il disegno dell'indagine prevede una metodologia di campionamento per quote[239] ed interviste esclusivamente dirette, senza rispondenti *proxy*. La decisione di utilizzare il metodo Cati per la rilevazione è stata determinata principalmente in virtù dell'abbattimento dei costi rispetto a strategie alternative; ciò ha comportato necessariamente una serie di problemi di carattere metodologico (contenimento della durata dell'intervista, formulazione semplice e d'immediata comprensione delle domande, riduzione del livello di strutturazione del questionario, aumento del rischio d'auto-selezione del campione). Tuttavia, una volta considerati i vantaggi e le criticità, il metodo Cati è risultato compatibile con gli obiettivi dell'indagine, in merito principalmente alla dimensione minima del campione necessaria per ottenere stime attendibili. Il piano metodologico del progetto è stato orientato alla riduzione degli effetti negativi che derivano da un'indagine telefonica; è stata prestata particolare attenzione alla dimensione del questio-

239 Per maggior ragguagli si legga il par. 9.2 relativo alla metodologia campionaria usata e ai parametri di stima.

nario, alla formulazione delle domande e alla classificazione delle risposte. Il livello degli errori non campionari è stato ridotto tramite una procedura apposita di trattamento dei contatti con un protocollo di gestione del numero telefonico estratto così sintetizzabile: 7 tentativi, in 2 giorni diversi (feriale e festivo) e in 3 orari diversi (11.30-15.00; 17.30-19.00; 19.00-23.00).

L'impianto copre interamente le fasi della vita lavorativa tipiche dell'individuo medio, con un approccio di tipo *life cicle*; tuttavia il mercato del lavoro attuale è sensibilissimo a variazioni congiunturali, alla tecnologia e alle normative, che restringono a pochi mesi l'arco di tempo necessario al verificarsi di evoluzioni (o involuzioni) occupazionali. Pertanto abbiamo previsto l'integrazione delle analisi trasversali (riferite all'anno) con una lettura longitudinale (riferita a più anni) che permetta di seguire sia l'individuo medio nel tempo (analisi di coorte o pseudo panel o di pool dei target) che i singoli e specifici individui (panel individuale). La dimensione del panel 2005-06 sarà di circa il 60% del campione 2005, con una dimensione campionaria pari a circa 23.000 interviste individuali ripetute.

Il servizio di raccolta dei dati è stato assegnato con bando di gara con procedura aperta. Nella fase di valutazione dei progetti operativi sono stati privilegiati gli aspetti che garantissero un elevato livello di qualità dei dati rilevati: numero di intervistatori dedicati all'indagine, formazione diretta degli intervistatori da parte dei ricercatori Isfol, limitato turnover del *call center*, tecniche di riduzione delle mancate risposte totali, monitoraggio continuo dei tassi di risposta. In tal modo è stato possibile coniugare una consistente numerosità del campione con un buon livello di qualità dei dati raccolti.

Con la società di rilevazione, la Doxa Spa, si è tenuto un rapporto di strettissima collaborazione e monitoraggio, oltre che di diretta formazione degli intervistatori. Il turnover del *call center* è stato molto contenuto e i tempi dell'indagine limitati a soli 3 mesi, tra il 15 gennaio e il 30 aprile 2005. La durata media dell'intervista è stata di 17 minuti, con un tasso di abbandono dell'intervista del 5%.

Per assicurare i fruitori delle analisi e dei microdati circa la metodologia applicata e le strategie adottate per contenere sia gli errori campionari che non campionari rimandiamo ai paragrafi seguenti, interamente dedicati alla metodologia campionaria. Ci preme tuttavia sottolineare come questa rilevazione sia stata ispirata e gestita con criteri nuovi, con una concezione della ricerca anche nella fase di costruzione e di dettaglio dell'indagine, in cui non esistono scatole chiuse, informazioni non disponibili alla cessione del dato; hanno partecipato al processo esperti delle singole materie presenti all'interno dell'Istituto, valorizzando le specializzazioni esistenti, e all'esterno, attraverso il coinvolgimento diretto di molti accademici e ricercatori pubblici, cui si è chiesto di collaborare nel definire le strategie e nel proporre soluzioni a nodi che, inevitabilmente, si creano in processi così complessi.

La validazione del questionario, del campionamento e dei dati è avvenuta, pertanto, attraverso alcuni semiari pubblici, alcuni incontri interni all'Istituto e attra-

verso una molteplice serie di incontri su temi specifici con esperti di primissimo piano, sia del mondo della ricerca[240] che istituzionale. Tutto ciò è stato realizzato dal gruppo di lavoro che coordina l'Indagine nell'intento di avere il massimo grado di partecipazione e di minimizzare eventuali errori metodologici.

240 L'indagine è stata presentata in 2 tavoli tecnici con esperti provenienti dal mondo accademico ed istituzionale:

- Roma, Venerdì 11 novembre 2005 La presentazione ha riguardato: impianto e questionario, campionamento, modulo donne, modulo giovani, modulo over 50, monitoraggio Legge 30, canali di ricerca e servizi pubblici per l'impiego, la formazione e l'istruzione (parte anagrafica). Relatori: Emiliano Mandrone, Marco Centra, Debora Radicchia, Emiliano Rustichelli, Valentina Meliciani, Massimiliano Tancioni. Discussant: Franco Frigo (Formazione continua, Isfol), Antonio Golini (Facoltà di Scienze statistiche, Università di Roma "la Sapienza") e Franco Peracchi (Facoltà di Economia, Università di Roma "Tor Vergata"), Lea Battistoni (Direttore generale Ministero del lavoro, Dipartimento mercato del lavoro). Sono Intervenuti, tra gli altri: Diana Gilli (Ricerche sui sistemi del lavoro, Isfol), Paolo Severati (Struttura Nazionale di Valutazione, Isfol), Marengon Maurizio (Regione Emilia Romagna), Roberto Torrini (Banca d'Italia).
- Roma, Venerdì 18 novembre 2005 L'incontro ha analizzato nel dettaglio i singoli quesiti della Plus05 e le migliorie possibili per la Plus06, questioni definitorie e possibili ulteriori analisi, anche panel. Relatori: Emiliano Mandrone, Debora Radicchia e Stefano Laj. Discussant: Daniele Checchi (Facoltà di Economia, Università di Milano), Riccardo Gatto (Forze lavoro, Istat), Pietro Cipollone (Banca d'Italia) e Franco Peracchi (Facoltà di Economia, Università di Roma "Tor Vergata"). Sono Intervenuti, tra gli altri: Manuel Marocco (questioni giuridiche), Roberto Landi (servizi pubblici per l'impiego), Germana di Domenico (servizi privati per l'impiego), Diana Gilli (Ricerche sui sistemi del lavoro, Isfol) e Dario Ercolani (telelavoro) dell'Isfol.
  Molteplici incontri, sia nella fase di disegno dell'Indagine che di analisi dei risultati, sono stati tenuti dal coordinamento dell'indagine con le aree dell'Isfol, tra cui Ricerche sui sistemi del lavoro, Analisi e valutazione delle politiche per il lavoro, Formazione continua, Formazione permanente, Struttura nazionale di valutazione, Fabbisogni formativi, con Carlo Dell'Aringa, allora presidente dell'Isfol.

9.2 IL PIANO DI CAMPIONAMENTO*

Il progetto prevede un'indagine campionaria riferita alla popolazione italiana residente in famiglia di età compresa fra i 15 e i 64 anni, con un'analisi approfondita su alcuni temi che si traducono nella definizione di particolari sottopopolazioni. Ne segue una struttura articolata secondo i seguenti target:

- Giovani (popolazione in età compresa tra 15 e 29 anni)
- Donne (popolazione femminile in età compresa tra 20 e 49 anni)
- Over 50 (popolazione in età compresa tra 50 e 64 anni)
- Disoccupati/inoccupati (persone in cerca di occupazione)
- Occupati (persone occupate)

L'obiettivo del piano di campionamento è stato dunque quello di produrre stime attendibili per l'intera popolazione oggetto di studio e per sottoinsiemi di essa definiti dai target sopra indicati.

I target, definiti come sottogruppi della popolazione di riferimento, sono parzialmente sovrapposti. Tale elemento determina una maggiore efficienza dell'indagine rispetto ad una strategia che preveda una singola *survey* per ogni sottopopolazione. Inoltre, ciascuna unità campionaria fa parte contemporaneamente di più di un target e contribuisce in tal modo a ridurre la numerosità campionaria totale. La sovrapposizione parziale delle sottopopolazioni aumenta quindi l'efficienza dell'indagine rispetto ai costi, una volta fissato il livello di attendibilità delle stime. Tuttavia le sottopopolazioni che definiscono i target non presentano un livello eccessivo di sovrapposizione, elemento questo che assicura che a ciascun individuo verranno proposti non più di 3 moduli del questionario, con il risultato di ridurre i tempi dell'intervista, aspetto particolarmente critico in una rilevazione telefonica, in ultima analisi, di diminuire l'incidenza degli errori non campionari.

Va precisato, circa la definizione delle sottopopolazioni riportata nello schema precedente, che la definizione della sottopopolazione Disoccupati/inoccupati si riferisce non solo alla versione più rigorosa dettata dall'Ilo (International Labour Organization) di *persone in cerca di occupazione*, ma anche ad una definizione allargata e comprensiva delle persone *inattive in cerca di occupazione.*

I target relativi ai giovani, alle donne e agli anziani sono stati ulteriormente ripartiti tra attivi e inattivi, in modo da rendere il campione in grado di fornire stime attendibili per:

- Giovani occupati (in età compresa tra 15 e 29 anni)
- Giovani studenti (in età compresa tra 15 e 29 anni)
- Giovani altra condizione (in età compresa tra 15 e 29 anni)

* Di Centra M., Isfol, Area "Analisi e valutazione delle politiche per l'occupazione", Falorsi P. D., Istat e Laj S., Isfol, Area "Ricerche sui sistemi del lavoro".

- Donne attive (in età compresa tra 20 e 49 anni)
- Donne inattive (in età compresa tra 20 e 39 anni)
- Anziani attivi (in età compresa tra 50 e 64 anni)
- Anziani inattivi (in età compresa tra 50 e 64 anni)

Il risultato è stato quello di portare da 5 a 9 le sottopopolazioni su cui produrre stime attendibili.
Per ottimizzare l'efficienza del piano di campionamento, sfruttando al meglio le conoscenze a priori sulla popolazione e per esigenze dettate dalla struttura del progetto dell'indagine che, oltre a voler fornire stime attendibili per ciascuno dei target definiti in precedenza, necessita di stime per alcuni parametri di tali sottopopolazioni disaggregati per domini di studio, si è optato per un piano di campionamento stratificato in cui gli strati sono determinati dalle variabili definite nella tab. 9.2 e dalle variabili strutturali che definiscono i domini di studio (localizzazione, età, genere, ecc.).
Il campione è stato quindi pianificato in modo da fornire stime attendibili per ciascuna sottopopolazione d'interesse, disaggregata per domini di studio, che includono il genere, la classe di età e la regione di residenza.
A tal fine è stata praticata una stratificazione della popolazione eleggibile in modo da riprodurre le sottopopolazioni definite in precedenza nei singoli domini di studio come aggregazione univoca di strati disgiunti.
Il piano di campionamento ha quindi definito un'allocazione del campione, di dimensione pari a 40.000 interviste, tale da fornire stime significative per i domini di studio pianificati *ex-ante* durante la fase di programmazione dell'indagine.
La pianificazione dei domini ha previsto la significatività delle stime per tutte le variabili di strato e quindi sia a livello territoriale: per regione e per gli 11 comuni italiani con popolazione superiore a 250.000 abitanti; che per genere, classi di età e condizione occupazionale.
La procedura per allocare il campione negli strati, vincolandola a produrre stime di attendibilità predefinita, presenta numerosi elementi di complessità. Tali elementi sono dovuti principalmente al fatto che la numerosità del campione in ciascuno strato contribuisce alla numerosità, e quindi all'attendibilità delle stime, di più sottopopolazioni.
È stato necessario utilizzare una procedura di allocazione multidominio, sviluppata appositamente per il progetto Plus, attraverso la soluzione di un sistema di allocazione vincolata ai livelli di varianza predefiniti:
sia data una popolazione stratificata in H strati U_h (h=1,2,…,H) di numerosità N_h; siano definiti D sottogruppi di popolazione, ciascuno disaggregato per domini di studio, determinati in modo che la numerosità della popolazione in ciascun dominio sia ottenuta tramite aggregazione di strati disgiunti; sia N_d (d=1,2,…,D) la numerosità del generico dominio nella popolazione. Allora:

$$N_d = \sum_{h=1}^{H} N_h \cdot I_{h,d}$$

dove $I_{h,d}$ indica se lo strato h appartiene al dominio d:

$$I_{h,d} = \begin{cases} 1 \ \ se \ \ h \in d \\ 0 \ \ se \ h \notin d \end{cases}$$

La varianza campionaria della stima di una frazione P della popolazione nel dominio d è data da:

$$V_d = \sum_{h=1}^{H} \frac{S_h^2 \cdot (N_h - n_h)}{(N_h - 1) \cdot n_h} \cdot \frac{N_h^2}{N^2} \cdot I_{h,d} \cong \sum_{h=1}^{H} \frac{S_h^2 \cdot (N_h - n_h)}{N_h \cdot n_h} \cdot \frac{N_h^2}{N^2} \cdot I_{h,d} =$$

$$= \sum_{h=1}^{H} \frac{S_h^2 \cdot N_h^2}{N^2 \cdot n_h} \cdot I_{h,d} - \sum_{h=1}^{H} \frac{S_h^2 \cdot N_h}{N^2} \cdot I_{h,d}$$

vale a dire che la quantità $V_d\,(p)$ può essere scomposta in due addendi dei quali uno dipende dalle quantità n_h, ovvero dall'allocazione del campione negli strati, mentre l'altro è indipendente dall'allocazione del campione ed è funzione della partizione della popolazione nei domini.
Posto:

$$V_{d0} = -\sum_{h=1}^{H} \frac{S_h^2 \cdot N_h}{N^2} \cdot I_{h,d}\,; \quad V_{d1} = \sum_{h=1}^{H} \frac{S_h^2 \cdot N_h^2}{N^2 \cdot n_h} \cdot I_{h,d}\,; \quad V_{dh}^2 = \frac{S_h^2 \cdot N_h^2}{N^2} \cdot I_{h,d}$$

si ottiene:

$$V_d = V_{d0} + V_{d1} \Rightarrow V_{d1} = V_d - V_{d0} \Rightarrow \sum_{h=1}^{H} \frac{V_{dh}^2}{n_h} = V_d - V_{d0} \qquad (1)$$

L'obiettivo è quello di ottenere un'allocazione del campione negli strati, ovvero un n_h, tale che si abbia una varianza minore o uguale di un certo valore stabilito *ex-ante*. Secondo il teorema di Kuhn-Tucker, specificando un V^*_d, limite superiore della varianza per ogni dominio e quindi tale che $V_d \leq V^*_d$, esiste un λ_d tale che:

$$n_h = \sqrt{\sum_{d=1}^{D} \ddot{e}_d \cdot V_{dh}^2} \qquad (2)$$

che ci fornisce l'allocazione desiderata.

La determinazione dei valori λ_d ha richiesto la messa a punto di un processo iterativo in grado di convergere alla soluzione. Di seguito sono riportati i passi dell'algoritmo di iterazione:
dato un valore del parametro $_{k-1}\lambda_d$ al passo k-1 tale valore viene aggiornato come segue:
1 si calcola $_k n_h$ tramite la (2)
2 si calcola $_k V_d$ tramite la (1)
3 si calcola il nuovo $_k \ddot{e}_d$ come:

$$_k \ddot{e}_d = {}_{k-1}\ddot{e}_d \cdot V_{dh}^2 \cdot \frac{({}_k V_d - V_{d0})^2}{(V_d^* - V_{d0})^2}$$

Fissando il primo valore del parametro $_0\ddot{e}_d = 1 \quad \forall\, d$ l'algoritmo risulta definito.
I passi 1, 2 e 3 vengono ripetuti fino a quando $| {}_{k+1}\ddot{e}_d - {}_k\ddot{e}_d | \leq \varepsilon$, piccolo a piacere.
Il risultato finale sarà dunque un'allocazione del campione negli strati n_h tale da soddisfare le condizioni di varianza massima imposti da V^*_d per ciascun dominio.
I vincoli imposti si riferiscono alle stime che il campione è chiamato a produrre per ciascuno dei target elencati in precedenza, in un insieme di domini di studio; i vincoli utilizzati sono riportati sinteticamente nella tabella seguente.

Tabella 9.1 Vincoli di varianza imposti nel piano di campionamento

Tipo dominio	p	cv	Numero di vincoli
Nazionali	0,01	0,15	9
Nazionali per sesso ed età	0,05	0,15	33
Regionali	0,10	0,19	171
Comuni metropolitani	0,10	0,20	99
Totale			**312**

Fonte: Isfol Plus 2005

Dunque nel caso di una delle nove sottopopolazioni definite in precedenza disaggregata per sesso o per classi di età avremo, per una stima pari al 5%, un *CV* del 15%.
Tutte le procedure descritte finora sono dirette a sfruttare al meglio le informazioni disponibili a priori sulla popolazione oggetto di studio. Nella fase di disegno tali informazioni hanno permesso di allocare il campione negli strati predeterminando il livello di affidabilità delle stime. Secondo la teoria inferenziale classica è necessaria, nella fase di estrazione del campione, la conoscenza per ogni singola unità statistica delle medesime informazioni utilizzate nella fase di pianificazione. Come noto non sono disponibili per l'intera popolazione italiana liste con un bagaglio informativo così elevato. L'unica alternativa percorribile rimane dunque una strategia che preveda un campione non probabilistico. L'assenza di liste di popolazione ha impedito infatti l'attribuzione di una probabilità di inclusione, fondamento della statistica inferenziale, a ciascun individuo negli strati della popolazione. La scelta è stata quindi orientata verso il campionamento per quote,

dove la numerosità di ciascuna quota ha coinciso con quella degli strati. L'impossibilità di conoscere la probabilità di inclusione e il parallelo ricorso al campionamento per quote, espone, com'è noto, l'intera rilevazione al rischio di distorsioni sensibili dovute all'auto-selezione del campione.

Tuttavia la grande quantità di informazione ausiliaria disponibile nella fase di disegno del campione ha permesso, da un lato, di controllare *ex-ante* il livello della distorsione, e, dall'altro, ha fornito la base per la messa a punto di un robusto piano di controllo e correzione a valle della raccolta dei dati. È stato possibile infatti contenere e correggere buona parte degli errori non campionari attraverso un lavoro dedicato alla scelta accurata delle variabili di stratificazione e nella dettagliata costruzione degli strati (980 strati); il controllo a posteriori inoltre, sfruttando ulteriormente le informazioni sulla popolazione di riferimento attraverso l'applicazione di metodi di calibrazione, ha permesso di riequilibrare le stime distorte con l'ausilio di totali noti della popolazione, come si vedrà in maniera più dettagliata nel paragrafo successivo. La scelta della strategia campionaria per quote è stata quindi assunta con la convinzione di poter contenere gran parte della distorsione osservabile, di poterla individuare ed infine correggere, grazie all'ampia disponibilità di informazione ausiliaria.

L'intera operazione di disegno del campione è stata possibile utilizzando i dati dalla Rilevazione trimestrale sulle forze di lavoro dell'Istat (media annuale del 2003). La grande mole di informazioni ricavate dalla Rtfl ha permesso quindi l'identificazione della popolazione eleggibile, la partizione di tale popolazione nei target oggetto di interesse, la disaggregazione nei domini di studio, ed infine, la stratificazione.

Di seguito sono riportate le caratteristiche che definiscono sia i target che i domini di analisi sui quali è stata pianificata, per ciascuno dei target, l'attendibilità delle stime. La stratificazione della popolazione coincide quindi con la distribuzione congiunta del collettivo secondo le caratteristiche riportate nella tab. 9.2.

Tabella 9.2
Variabili di stratificazione

Regione	1	Piemonte e Valle d'Aosta
	3	Lombardia
	4	Trentino A.A.
	5	Veneto
	6	Friuli V.G.
	7	Liguria
	8	Emilia Romagna
	9	Toscana
	10	Umbria
	11	Marche
	12	Lazio
	13	Abruzzo
	14	Molise
	15	Campania
	16	Puglia
	17	Basilicata
	18	Calabria
	19	Sicilia
	20	Sardegna
Tipo comune	1	Comune metropolitano
	2	Comune non metropolitano
Genere	1	Maschi
	2	Femmine
Età in classi	1	15-19
	2	20-29
	3	30-39
	4	40-49
	5	50-64
Condizione occupazionale	1	Occupato
	2	In cerca di occupazione
	3	Studente
	4	Pensionato
	5	Altro inattivo

Fonte: Isfol Plus 2005.

9.3 RIPORTO ALL'UNIVERSO E STIMATORE DI CALIBRAZIONE*

Nella fase successiva alla rilevazione si è proceduto nel calcolo dello stimatore o coefficiente di riporto all'universo. La strategia utilizzata non ha permesso la costruzione di uno stimatore secondo le usuali tecniche inferenziali[241]. È stato seguito perciò un approccio predittivo basato su modelli di superpopolazione[242]. Avendo a disposizione i totali noti per una serie di variabili esplicative derivati dalla popolazione di riferimento (Rilevazione trimestrale delle forze di lavoro - Istat) si è deciso di utilizzare lo stimatore di regressione basato su variabili strumentali[243]. Tale stimatore, oltre a sfruttare le informazioni delle variabili ausiliare, gode di una serie di proprietà tra le quali quella della calibrazione, secondo la quale le stime dei totali delle variabili ausiliarie utilizzate come regressori, corrispondono ai totali noti. In questo modo è possibile quindi calibrare la popolazione stimata secondo i totali noti della popolazione reale per alcune variabili e correggere eventuali distorsioni nella stima della popolazione sfruttando così al meglio le informazione già note sulla popolazione.
Si definisca con U una popolazione di N elementi frazionata in h strati (h=1,…,H) di dimensione N_h e si indichi con k (k=1,…,N) il generico elemento di essa, con riferimento al quale si denotino con y_k, $\mathbf{x}_k$ e $\mathbf{z}_k$ e rispettivamente il valore della variabile d'interesse, di un vettore di variabili ausiliarie e di un vettore di variabili strumentali.
Si supponga di osservare un campione s di n elementi, essendo $s_h = s \cap U_h$ il campione di dimensione n_h osservato nello strato h-esimo.
Per stimare il generico parametro d'interesse

$$Y = \sum_U y_k$$

si dispone della seguente situazione informativa:

- per le unità appartenenti al campione sono noti i valori delle variabili y_k, $\mathbf{x}_k$ e $\mathbf{z}$;
- è conosciuto il totale

$$\mathbf{X}_U = \sum_U \mathbf{x}_k$$

  delle variabili ausiliarie; per la condizione precedente è quindi noto il valore del vettore

---

* Di Centra M., Isfol, Area "Analisi e valutazione delle politiche per l'occupazione", Falorsi P. D., Istat e Laj S., Isfol, Area "Ricerche sui sistemi del lavoro"; si ringrazia, inoltre, G. Linfante per l'aiuto nella realizzazione della calibrazione.

241 In particolare non è stato possibile costruire uno stimatore ottenuto secondo il disegno di Horvitz-Thompson. (Horvitz, D.G. e Thompson, D.J. - 1952).

242 Royall, R.M. (1970).

243 La procedura utilizzata fa riferimento alla metodologia degli stimatori di calibrazione di Särndal e Lundström (2005) e Deville e Särndal (1992).

$$\mathbf{X}_{\bar{s}} = \mathbf{X}_U - \mathbf{X}_s = \sum_U \mathbf{x}_k - \sum_s \mathbf{x}_k$$

relativo alla parte non osservata nel campione.

In un ottica predittiva, la stima del totale della variabile di interesse viene ottenuto mediante la somma di due parti distinte: il totale osservato sul campione

$$Y_s = \sum_s y_k$$

e il totale dei valori predetti $\tilde{y}_k$ della variabile di interesse

$$\tilde{Y}_{\bar{s}} = \sum_{\bar{s}} \tilde{y}_k \, .$$

Al fine di stimare $\tilde{y}_k$ si suppone l'esistenza di un modello lineare che lega la variabile d'interesse alla variabile ausiliaria,

$$y_k = \mathbf{B}'\mathbf{x}_k + \varepsilon_k$$

in cui $\mathbf{B}$ denota il vettore dei parametri della regressione ed ε_k il residuo casuale, i cui valori attesi sotto il modello di superpopolazione utilizzato sono dati da:

$$E(\varepsilon_k) = 0\,, \qquad E(\varepsilon_k)^2 = V(\varepsilon_k) = \acute{o}^2 \tag{3}$$

Mediante una tecnica di regressione fondata sull'uso delle variabili strumentali[244] il vettore $\mathbf{B}$ può essere stimato come

$$\hat{\mathbf{B}} = \left(\sum_s \mathbf{x}_k \mathbf{z}_k'\right)^{-1} \sum_s \mathbf{z}_k y_k \, .$$

In tal modo il valore predetto $\tilde{y}_k$ è dato da:

$$\tilde{y}_k = \hat{\mathbf{B}}'\mathbf{x}_k = \left[\left(\sum_s \mathbf{x}_k \mathbf{z}_k'\right)^{-1} \sum_s \mathbf{z}_k y_k\right] \mathbf{x}_k$$

Pertanto, lo stimatore del totale della variabile d'interesse può essere espresso come[245]:

$$\tilde{Y}_{REG} = Y_s + Y_{\bar{s}} = \sum_s y_k + \hat{\mathbf{B}}'\left(\sum_{\bar{s}} \mathbf{x}_k\right), \tag{4}$$

244 Johnston, J. (1984), Econometric Methods.

245 *Sistemi di stima generalizzata*, di Centra M. e Falorsi P.D, Temi & Strumenti, Isfol.

9.3 Riporto all'universo e stimatore di calibrazione

Nello specifico, per le unità appartenenti al generico strato U_k, il vettore di variabili strumentali è definito moltiplicando il vettore di variabili ausiliarie per il reciproco del tasso di sondaggio nello strato, denominato anche come *peso base*.

$$\mathbf{z}_k = \mathbf{x}_k a_k \qquad \text{per} \qquad k \in U_h$$

essendo

$$a_k = \frac{N_h}{n_h}.$$

Mediante semplici passaggi la (4) può essere espressa in forma lineare come

$$\tilde{Y}_{\text{REG}} = \sum_{\text{s}} w_k \, y_k \tag{5}$$

essendo

$$w_k = a_k \, g_k$$

in cui

$$g_k = 1 + \sum_{\bar{s}} x_k \left(\sum_s x_k z_k \right)^{-1} z_k \, .$$

Lo stimatore adottato è calibrato in quanto esso garantisce la condizione

$$\tilde{\mathbf{X}}_{REG} = \sum_s \mathbf{x}_k \; w_k = \mathbf{X}_U$$

Perché la calibrazione porti ad un significativo miglioramento dell'attendibilità delle stime, le variabili ausiliarie devono essere correlate con la variabile oggetto di studio. Più la correlazione è alta e più forte sarà il guadagno di efficienza, al contrario nel caso in cui non ci sia alcun legame tra le variabili tale procedura potrebbe portare ad una perdita di efficienza delle stime.
Dunque dopo un'analisi accurata delle variabili da utilizzare come regressori, basata sia su eventuali distorsioni tra parametri noti della popolazione e corrispondenti stime campionarie, sia sulle correlazioni corrispondenti, si sono individuate le variabili riportate nella tab. 9.3.

Tipo di attività	1	Dipendente a tempo indeterminato
	2	Dipendente a tempo determinato
	3	Autonomo
Part-time	1	Part-time
	2	Full-time
Titolo di studio	1	Licenza elementare
	2	Licenza media
	3	Diploma
	4	Laurea

Tabella 9.3 Variabili ausiliarie utilizzate per la calibrazione

Fonte: Isfol Plus 2005

Ognuno di questi totali è stato poi ulteriormente disaggregato secondo il genere, le classi di età e l'area geografica, generando 202 vincoli. Il risultato finale di tale procedura ha restituito un correttore moltiplicativo del *peso base* tale che le stime della popolazione, per le caratteristiche sopra indicate, risultino coincidenti a quelle note della popolazione, pur mantenendo i vincoli imposti nel calcolo dei coefficienti di riporto all'universo, ovvero i totali noti per gli strati di campionamento definiti dalle variabili della tab. 9.2.

9.4 ANALISI DELLA VARIANZA DELLE STIME*

L'errore di stima è dato dalla distanza tra il valore stimato e quello reale, ovvero:

$$\left(\tilde{Y}_{REG} - Y\right) = \sum_s y_k w_k - \left(\sum_s y_k + \sum_{\bar{s}} y_k\right) =$$

$$= \sum_s y_k \left(w_k - 1\right) - \sum_{\bar{s}} y_k$$

Per la (3) si ha:

$$V\left(\tilde{Y}_{REG} - Y\right) = \sum_s \left(w_k - 1\right)^2 \sigma^2 + \sum_{\bar{s}} \sigma^2$$

$$= \sigma^2 \left[\sum_s \left(w_k - 1\right)^2 + \left(N - n\right)\right].$$

Essendo,

$$\sum_s w_k = N$$

la precedente espressione può essere riformulata come:

$$V\left(\tilde{Y}_{REG} - Y\right) = \sigma^2 \left[\sum_s \left(w_k - 1\right)^2 + \sum_s \left(w_k - 1\right)\right] =$$

$$= \sigma^2 \left[\sum_s \left(w_k^2 + 1 - 2w_k + w_k - 1\right)\right] =$$

$$= \sigma^2 \left[\sum_s \left(w_k^2 - w_k\right)\right] = \sigma^2 \left[\sum_s w_k \left(w_k - 1\right)\right].$$

Per ogni singola unità *k,* una stima robusta della varianza σ^2 è data dal valore quadratico del residuo stimato

$$\tilde{\varepsilon}_k = y_k - \tilde{y}_k .$$

Pertanto, una stima robusta della varianza da modello è ottenibile, mediante una tecnica di tipo *plug-in* come:

$$\tilde{V}\left(\tilde{Y}_{REG} - Y\right) = \sum_s w_k \left(w_k - 1\right) \tilde{\varepsilon}_k^2 \qquad (6)$$

* Di Centra M., Isfol Area "Analisi e valutazione delle politiche per l'occupazione", Falorsi P. D., Istat e Laj S., Isfol Area "Ricerche sui sistemi del lavoro".

Attraverso questa ultima espressione si sono calcolati i livelli di significatività per alcune stime. Come misura dell'attendibilità delle stime stesse è stato scelto il coefficiente di variazione definito come:

$$\tilde{C}V(\tilde{Y}_{REG}) = \frac{\sqrt{\tilde{V}(\tilde{Y}_{REG})}}{\tilde{Y}_{REG}}$$

Non potendo pubblicare, per ogni dominio, il relativo valore del CV per tutte le stime prodotte dall'indagine, si è proceduto alla modellizzazione di tali valori in modo da poter fornire uno schema sintetico dei livelli di significatività delle stime e gli strumenti, per chi lo desiderasse, per poter effettuare il calcolo autonomamente per qualunque tipo di stima.
La nuvola di punti descritta dalla stima ed il relativo coefficiente di variazione viene successivamente interpolata in modo da poter ricavare, per ogni stima prodotta dall'indagine, il relativo livello di significatività. Il modello utilizzato è del tipo seguente:

$$\log \tilde{C}V\left[(\tilde{Y}_{REG})^2\right] = b_0 + b_1 \log(\tilde{Y}_{REG}). \qquad (7)$$

I parametri del modello (7) sono stati stimati per le stime a livello nazionale e separatamente per regione.
Sulla base della conoscenza dei valori stimati $\tilde{b}_0$ e $\tilde{b}_1$ dei parametri del modello (7), è possibile ricavare il CV di una generica stima $\tilde{Y}_{REG}$ mediante l'espressione:

$$\tilde{C}V(\tilde{Y}_{REG}) = \sqrt{\exp\left[\tilde{b}_0 + \tilde{b}_1 \log(\tilde{Y}_{REG})\right]} \qquad (8)$$

Nella tab. 9.4 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 CV di ciascuna stima pubblicata. Ad esempio per una stima di 50.000 nella regione Liguria, riportando i coefficienti della tab. 9.4 relativi alla Liguria nell'espressione (8) è possibile determinare il valore del CV come:

$$\tilde{C}V(50.000) = \sqrt{\exp\left[4{,}82251 + (-0{,}80168)\log(50.000)\right]} = 0{,}146$$

Tabella 9.4 Parametri del modello per il calcolo dell'attendibilità delle stime per domini

Dominio	Parametri del modello		
	B0	B1	R2
Nazionale	5,33823	-0,84086	0,83996
Regione			
Valle d'Aosta, Piemonte	10,01062	-1,28679	0,89227
Liguria	4,82251	-0,80168	0,7523
Lombardia	6,09309	-0,86157	0,72769
Trentino Alto Adige	7,08273	-1,08424	0,86287
Veneto	6,92977	-0,96176	0,86821
Friuli Venezia Giulia	5,00882	-0,82912	0,76995
Emilia Romagna	5,08807	-0,77926	0,77077
Toscana	7,82855	-1,05639	0,88272
Marche	6,97739	-1,04428	0,74876
Umbria	5,77504	-0,93596	0,83223
Lazio	5,65057	-0,84926	0,87675
Molise	5,52212	-0,9987	0,87357
Abruzzo	4,75082	-0,78541	0,69983
Campania	6,23804	-0,89488	0,85775
Puglia	4,99193	-0,7646	0,81007
Basilicata	5,84362	-0,99685	0,89087
Calabria	5,62878	-0,85916	0,80749
Sicilia	5,3623	-0,81192	0,82721
Sardegna	6,22144	-0,95926	0,88821

Fonte: Isfol Plus 2005

Nella tab. 9.5 si forniscono invece i livelli delle stime, per ciascun dominio, corrispondenti a determinati valori del CV. Generalmente per valori del CV minori di 0,15 la stima si considera *affidabile*, mentre se il CV è superiore a 0,25 la stima è da considerarsi *non affidabile*. Chiaramente per valori intermedi l'affidabilità della stima è definita *critica*. Riprendendo l'esempio precedente ad una stima di 46.557 nella Liguria corrisponde un livello di significatività di 0,15. Per la stessa regione la stima di 50.000 forniva un CV compreso tra il 0,10 ed 0,15, ovvero pari a 0,146.

Tabella 9.5 Livelli delle stime per determinati valori del CV per domini

Dominio	Livello di significatività (valori del CV)			
	0,1	0,15	0,2	0,25
Nazionale	136.656	52.095	26.280	15.457
Regione				
Valle d'Aosta, Piemonte	85.675	45.621	29.173	20.623
Liguria	128.023	46.557	22.714	13.017
Lombardia	246.995	96.365	49.419	29.440
Trentino Alto Adige	48.040	22.740	13.376	8.863
Veneto	161.707	69.590	38.259	24.055
Friuli Venezia Giulia	108.604	40.839	20.403	11.911
Emilia Romagna	252.457	89.175	42.617	24.036
Toscana	129.314	60.015	34.812	22.817
Marche	65.606	30.178	17.394	11.345
Umbria	65.539	27.556	14.902	9.251
Lazio	175.631	67.594	34.330	20.298
Molise	25.348	11.254	6.326	4.046
Abruzzo	149.077	53.089	25.519	14.457
Campania	182.941	73.919	38.862	23.602
Puglia	282.619	97.856	46.108	25.721
Basilicata	35.662	15.809	8.876	5.673
Calabria	148.981	57.972	29.674	17.652
Sicilia	214.586	79.038	38.911	22.457
Sardegna	79.730	34.236	18.793	11.802

Fonte: Isfol Plus 2005

Nella costruzione del campione i vincoli imposti sulla significatività delle stime, ovvero quei vincoli tali da ottenere una numerosità minima per strato che garantisca un livello pre-definito di significatività delle stime, non sono stati calcolati sui parametri stessi ma bensì sulle proporzioni, ossia sul peso che il parametro ha all'interno dello strato. Questo spiega la forte variabilità all'interno delle colonne della tab. 9.5, ovvero la variabilità dei livelli delle stime all'interno dei domini che si riferiscono ad un medesimo grado di significatività.
La stima della varianza campionaria ci fornisce anche la possibilità di calcolare un intervallo di confidenza per la stima stessa e dunque un spazio entro il quale il parametro è contenuto con una certa probabilità. Al livello di confidenza pari al 95%, i limiti inferiori e superiori di tale intervallo sono dati rispettivamente da:

$$\tilde{Y}_{REG} - 1{,}96\tilde{Y}_{REG}\,\mathrm{CV}(\tilde{Y}_{REG}) \;\; ; \;\; \tilde{Y}_{REG} + 1{,}96\tilde{Y}_{REG}\,\mathrm{CV}(\tilde{Y}_{REG}) \; .$$

9.5 CORREZIONE ED IMPUTAZIONE DELLE MANCATE RISPOSTE*

9.5.1 Individuazione ed eliminazione degli *outliers*

L'Indagine Plus rileva i redditi da lavoro chiedendo agli occupati la propria retribuzione. Con il fine generale di favorire l'intervistato a fornire l'informazione richiesta e di preservare la qualità del dato si sono seguiti diversi accorgimenti. L'utilizzo del sistema Cati (Computer Assisted Telephone Interviewing) e dell'indagine pilota preliminare per testare il questionario hanno già ridotto notevolmente le eventuali possibilità di errori. Come anticipato già nel capitolo 6 la domanda sui redditi è stata distinta a seconda della diversa tipologia occupazionale differenziando tra dipendenti, autonomi e collaboratori (co.co.co, lavoro a progetto e collaborazione occasionale) in modo da ridurre al minimo i calcoli necessari all'intervistato per fornire la risposta[246]. Nella fase stessa della rilevazione è stato imposto ai rilevatori, tramite il sistema Cati, di dover digitare due volte il reddito dichiarato per limitare al minimo errori in fase di battitura e poter correggere eventuali errori in corso d'opera. Le variabili del reddito, oltre a subire i controlli di qualità per la validazione del database, sono sottoposte ad un processo ulteriore di validazione che si articola concretamente in due fasi. In primo luogo l'individuazione e l'eventuale eliminazione degli *outliers* e successivamente l'imputazione delle mancate risposte riproducendo anche i valori degli *outliers* eliminati. Si definiscono *outliers* quei valori che si collocano nelle code della distribuzione e che risultano sensibilmente diversi dalla media (o dalla mediana), ovvero quelle unità statistiche che risultano significativamente diverse da quelle della maggior parte delle unità. Tali valori, soprattutto quelli di grandezza elevata, possono sia influire significativamente ed in maniera distorsiva sulla distribuzione stessa del carattere osservato e conseguentemente sulle relative statistiche descrittive (quindi sui valori delle medie da stimare), sia sulla successiva stima dei valori mancati.

Per l'individuazione degli *outliers* si è utilizzato il metodo proposto da Hidiroglou e Berthelot che si basa sul ricorso alle soglie di accettazione ricavate dai quartili della distribuzione. In pratica consiste nell'applicare il metodo dei quartili ad una trasformata z della variabile iniziale y oggetto di studio.

L'intervallo di accettazione di una generica osservazione y è definito in base al valore della mediana e dei quartili della distribuzione come:

$$A_y = (Me_y - c_{inf}\, d_{inf,y};\ Me_y + c_{sup}\, d_{sup,y}) = (A_{inf,y}\ ;\ A_{sup,y})$$

* Di Laj S., Isfol, Area "Ricerche sui sistemi del lavoro".

246 Ai dipendenti è stato chiesto il reddito netto mensile, agli autonomi il reddito lordo annuo ed ai collaboratori il reddito lordo mensile.

dove $d_{inf,y}$ e $d_{sup,y}$ sono rispettivamente gli scarti interquartili inferiore e superiore ovvero:

$$d_{inf,y} = M_{ey} - Q_{1y}$$
$$d_{sup,y} = Q_{3y} - M_{ey}$$

mentre c_{inf} e c_{sup}[247] sono fattori arbitrari che nel caso siano entrambi posti uguali a 1 l'intervallo di accettazione è pari allo scarto interquartile.
Verranno quindi calcolate le soglie di accettazione alla trasformata della variabile y definita come:

$$z_y = \begin{cases} \dfrac{y - Me_y}{y} & se \quad 0 < y < Me_y \\ \dfrac{y - Me_y}{Me_y} & se \quad y \geq Me_y \end{cases}$$

.
Nell'Indagine Plus sono risultati fuori dall'intervallo di accettazione un totale di 30 casi che sono stati dunque eliminati e trattati successivamente come mancate risposte da imputare (tab. 9.6).

Tabella 9.6 Analisi degli outliers per le variabili del reddito

	Autonomi	Collaboratori	Dipendenti
N	851	403	9.730
Q1	12.000	500	900
Me	24.000	800	1.100
Q3	40.000	1.250	1.400
Ainf	585	32	111
Asup	664.000	18.800	13.100
y<Ainf	10	0	14
y>Asup	2	1	3
Totale outliers	**12**	**1**	**17**

Fonte: Isfol Plus 2005

Nel paragrafo successivo andremo ad analizzare la tecnica utilizzata per il trattamento delle mancate risposte per le variabili relative al reddito.

247 Nell'Indagine Plus si è utilizzato un valore tale da non rendere troppo vincolanti le soglie di accettazione e pari a c_{inf}=c_{sup}=40.

9.6 PROCEDURA DI IMPUTAZIONE DELLE MANCATE RISPOSTE*

Il presente paragrafo fornisce alcune informazioni sulla metodologia statistica utilizzata per l'imputazione dei valori mancanti relativi alle variabili reddito da lavoro dipendente (v658_1), reddito da lavoro autonomo (v650_1) e reddito da lavoro per i collaboratori (v655_1). Le serie affette da informazione incompleta sono state imputate ricorrendo a tecniche standard di tipo "donatore", con schemi di stratificazione iniziale. La scelta di adottare tali tecniche, in luogo di procedure più raffinate (es. *multiple imputation*), è stata dettata principalmente dalla natura del dato osservato. La ridotta dimensione del campione di "eleggibili" in alcuni casi (es. variabile v655_1) e l'alta percentuale di mancata risposta in altri (es. variabile v650_1) hanno reso impraticabile la sperimentazione di tecniche alternative e più raffinate. Inoltre, per garantire una certa omogeneità nella tecnica di imputazione si è preferito adottare la medesima procedura per tutte e tre le variabili che compongono il sottomodulo sul reddito da lavoro. Ciò è andato a discapito della variabile reddito da lavoro dipendente (variabile v658_1), la cui dimensione della mancata risposta avrebbe consentito di utilizzare tecniche più raffinate. Ne è scaturita pertanto, l'adozione di un'unica tecnica di imputazione per le tre variabili di riferimento, consistente nella procedura *hot-deck* versione *Approximate Bayesian Bootstrap* (Abb)[248].

La tecnica d'imputazione *hot-deck* è un metodo semi-parametrico che consente la ricostruzione dell'informazione mancante senza richiedere alcuna assunzione a priori sulla forma distributiva del dato, basandosi esclusivamente sulla conoscenza della distribuzione empirica del dato osservato. La preventiva stratificazione del campione secondo variabili assumibili come predittori del dato mancante, consente di raffinare ulteriormente la scelta del donatore nel *pool* di quelli potenziali.

Più formalmente, sia la matrice **Z** di dimensione $i \times n$ composta dei vettori colonna ($Z_1, Z_2, \ldots, Z_i$). Sia Z_k il vettore affetto da informazione incompleta. Indicheremo con $\mathbf{Z}_{com}$ la componente osservata del dato, di dimensione $n_1<n$ e con $\mathbf{Z}_{inc}$ la componente inosservata, di dimensione n_0, con $n = n_1 + n_0$. La fig. 9.1 fornisce una rappresentazione del data set **Z**, i rettangoli sulla sinistra indicano rispettivamente la componente del dato osservato Z_{com} e del dato incompleto Z_{inc}. Un modello di imputazione genera un data set completo estraendo dalla distribuzione a posteriori di z_{inc}.

* Di Tuzi F., Istat.

248 Cfr. Rubin and Schenker (1986) e Rubin (1987).

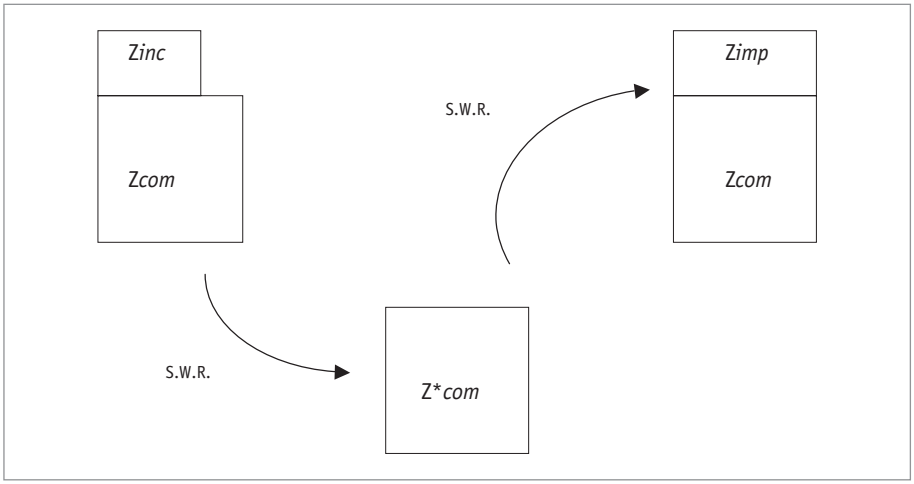

Figura 9.1
The Approximate Bayesian Bootstrap

Nella sua versione più semplice, le n_0 righe di Z_{inc} sono estratte in maniera casuale e con reinserimento dalla distribuzione dei dati noti Z_{com}. Al fine di garantire maggiore accuratezza nella scelta del dato da imputare, la procedura *hot-deck* nella versione ABB procede ad una prima estrazione con reinserimento di Z_{com} per generare Z^*_{com} (avente le stesse dimensioni di Z_{com}) ed in un secondo step, n_0 righe sono campionate con reinserimento dalla distribuzione di Z^*_{com} per generare il data set imputato. Il metodo può essere ulteriormente raffinato nel caso si sospetti una qualche variabilità tra strati del meccanismo generatore del dato mancante. In questo caso, si possono ottenere stime più accurate applicando il metodo *hot-deck* separatamente per ogni strato. Ovviamente, affinché l'estrazione garantisca le buone proprietà degli stimatori, sono richiesti requisiti minimi sulla numerosità delle unità all'interno degli strati. Prima di passare alla descrizione dei risultati dell'imputazione delle tre variabili è utile far rilevare alcune limitazioni della procedura *hot-deck*.

Come molte delle tecniche di imputazione tradizionale, la procedura *hot-deck* soffre del vincolo di "ignorabilità" del processo generatore del dato mancante. Nel caso in cui ipotesi di "ignorabilità" non fosse vera, l'analisi sul campione ridotto che non tenesse conto di questo fatto sarebbe soggetta a distorsione[249].

Formalmente, date due variabili X e Y, le quali possono essere o meno osservate, la probabilità di risposta R:

1 può dipendere da X ma non da Y
2 può essere indipendente da X e Y
3 può dipendere da X e Y.

249 Cfr. Little and Rubin (1987) sulle implicazioni statistiche della violazione delle ipotesi Mar e Mcar.

Nel primo caso si dice che i dati mancanti sono di tipo *missing at random* (*Mar*), nel senso che i valori di *Y* osservati non sono necessariamente un sottocampione casuale dei valori campionati, bensì un campione casuale dei valori campionati all'interno di una sottoclasse definita dai valori di *X.* Nel secondo caso, i dati mancanti sono *missing at random* (*Mar*), mentre i dati osservati sono *observed at random* (*Oar*), o più semplicemente tutto il meccanismo di dati mancanti è di tipo *missing completely at random* (*Mcar*). Nel terzo caso infine, i dati incompleti e osservati non sono né *Mar* né *Oar* e il meccanismo di dati mancanti è di tipo *non missing at random* (Nmar).
Più in generale, data la funzione di risposta

$$f(R|Y_{obs}, Y_{mis}, \xi)$$

il meccanismo generatore dei dati mancanti può risultare tale che:

a se non dipende dalla componente inosservata Y_{mis}, ovvero

$$P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, \xi)$$

allora il meccanismo generatore dei dati mancanti è di tipo *missing at random* (Mar);

b se non dipende né dalla componente inosservata Y_{mis}, né dalla componente osservata Y_{obs}, ovvero

$$P(R|Y_{obs}, Y_{mis}, \xi) = P(R|\xi)$$

allora il meccanismo sottostante la generazione del dato mancante è di tipo *missing completely at random* (Mcar);

c se dipende sia dalla componente inosservata Y_{mis}, che dalla componente osservata Y_{obs}, ovvero

$$P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, Y_{mis}\xi)$$

allora il meccanismo sottostante la generazione del dato mancante è di tipo *non missing at random* (Nmar).

Un meccanismo generatore del dato incompleto è definito "ignorabile" se i valori mancanti sono *missing at random*, nel senso che le unità osservate rappresentano un sottocampione casuale delle unità campionate. Se la probabilità che y_i sia osservata dipende dal valore di y_i, allora il meccanismo di dati mancanti non è ignorabile e l'analisi sul campione ridotto che non tenga in considerazione questo fatto è soggetto a distorsione. L'assunzione di ignorabilità di un processo, benché criticabile in alcune circostanze, porta spesso a risultati migliori rispetto all'adozione di procedure *ad hoc* (ad esempio quelle basate sull'utilizzo della sola

componente osservabile), in quanto consente di rimuovere tutta la distorsione generata dalla parte di non risposta spiegabile attraverso la componente osservata Y_{obs}. È utile tenere a mente che l'assunzione cruciale su cui si basano i metodi che ignorano il processo di mancata risposta, è che la propensione alla non risposta possa essere spiegata dai dati osservati e non che essa sia totalmente incorrelata con i dati mancanti. È per questo che procedure come la *hot-deck* usualmente producono buoni risultati anche quando l'ignorabilità è sospettata. In generale, la plausibilità dell'ipotesi di ignorabilità dipende dal contenuto dei dati mancanti e dalla complessità del modello dei dati $P(Y|\vartheta)$; se i dati osservati contengono informazione sufficiente per prevedere i dati mancanti e se il modello è sufficientemente elaborato da far uso di questa informazione, ci si aspetterà che la dipendenza residua di R da Y_{mis} dopo aver condizionato ad Y_{obs} sarà relativamente minore.

9.6.1 Applicazione all'indagine

La tab. 9.7 riporta il numero delle unità affette da informazione mancante sulle tre variabili d'interesse e la rispettiva incidenza sul sottocampione di riferimento:

Tabella 9.7 Valori assoluti e incidenza del dato mancante sulle variabili reddito

	Completi	% incidenza	Mancanti	% incidenza
v 650_1	851	34,0	1655	66,0
v 655_1	403	64,4	223	35,6
v 658_1	9730	76,4	3006	23,6

Fonte: Isfol Plus 2005.

Come anticipato in apertura, è necessario sottolineare due aspetti nelle distribuzioni delle mancate risposte che hanno condizionato la scelta della procedura di imputazione e che dovrebbero essere tenute a mente nel caso si decidesse di utilizzare le serie imputate dei dati. Primo, l'alta percentuale del dato mancante, in particolare per la variabile riferita al reddito da lavoro autonomo (v650_1), secondo, e di fondamentale rilevanza, la bassa incidenza delle unità potenzialmente utilizzabili per l'imputazione, sia per la variabile reddito da lavoro autonomo (v650_1) che per la variabile reddito da lavoro per i collaboratori (v655_1).
Al fine di migliorare la qualità della serie imputata, la procedura *hot-deck* è stata affiancata da uno schema di stratificazione iniziale. Il pool dei potenziali donatori è stato individuato ricorrendo ad un insieme di probabili predittori del processo generatore del dato mancante. È necessario precisare che il numero delle esplicative è stato scelto di volta in volta in funzione della variabile da imputare e quindi varia tra il reddito da lavoro autonomo (v650_1), il reddito da lavoro per i collaboratori (v655_1) e il reddito da lavoro dipendente (v658_1). Questa scelta è stata dettata primariamente da motivazioni di ordine statistico. Infatti, l'eccessiva stratificazione del pool dei donatori, individuati dalla combinazione delle

esplicative, in alcuni casi poteva dar luogo alla violazione della condizione sul numero minimo di donatori necessaria per garantire le buone proprietà degli stimatori[250].

Uno studio preliminare sulla struttura delle correlazioni tra variabili da imputare e predittori del meccanismo generatore del dato mancante, ha individuato le seguenti variabili di stratificazione: sesso, età, grado di istruzione e area regionale di appartenenza. A seconda della numerosità del dato da imputare e dei potenziali donatori, le variabili di stratificazione sono state opportunamente ricodificate. In particolare:

- per la variabile di selezione v658_1 (reddito da lavoro dipendente) sono state adottate le seguenti riclassificazioni, sesso (maschi, femmine), età in classi (15-29, 30-39, 40-49, 50-64), area geografica (Sud e Isole, Centro, Nord), livello di istruzione (licenza media, diploma, laurea e post-laurea);
- per la variabile di selezione V655_1 (reddito da lavoro per i collaboratori) sono state adottate le seguenti riclassificazioni, sesso (maschio, femmina), età in classi (meno di 30 e più di 30), area geografica (Sud e Isole, Centro-Nord), livello di istruzione (meno di diploma, più di diploma);
- per la variabile di selezione v650_1 (reddito da lavoro autonomo) sono state adottate le seguenti riclassificazioni, sesso (maschio, femmina),età in classi (meno di 30 e più di 30), area geografica (Sud e Isole, Centro-Nord), livello di istruzione (meno di diploma, più di diploma).

Malgrado l'aggregazione delle variabili di classificazione abbia indotto una riduzione delle dimensioni degli spazi di stratificazione, la condizione sul numero minimo di donatori è stata violata, anche se in un numero ridotto di casi. Del resto era prevedibile attendersi questo risultato, a causa del ristretto numero sia delle osservazioni sia del pool dei potenziali donatori. Per la v655_1 il programma ha segnalato la presenza di 1 strato con un solo dato osservato e di 5 strati con 2-5 dati osservati. Per la variabile v650_1, uno strato con una sola osservazione e uno strato con 2-5 dati osservati.

Analisi pre e post imputazione, hanno evidenziato come la serie ricostruita restituisca in generale una distribuzione simile a quella originaria. In particolare, la serie imputata per la v658_1 riproduce quasi esattamente la struttura distributiva del dato iniziale. Per le variabili v650_1 e v655_1, l'elevata percentuale del dato mancante e la scarsa numerosità del campione utilizzato per l'imputazione, non sembrano aver distorto la struttura distributiva del dato. Vi è tuttavia ragione di credere che l'alta percentuale di dati *missing* e la bassa numerosità dei potenziali donatori possa aver contribuito ad appiattire la distribuzione sui valori noti e ciò non sembra essere del tutto irrilevante, in un contesto in cui le categorie di lavoratori autonomi non sono del tutto omogenee rispetto ai redditi da lavoro.

250 Affinché siano rispettate le buone proprietà di uno stimatore, è necessario che il pool dei donatori sia non inferiore alle 5 unità.