Metodi per il trattamento delle diverse componenti della mancata risposta totale applicati all'indagine Istat sulla Disabilità[1]

Claudia De Vitiis[2], *Alessio Guandalini*[3], *Francesca Inglese*[4], *Marco D. Terribili*[5]

Sommario

Gli effetti negativi della mancata risposta totale sulle stime di un'indagine campionaria devono essere opportunamente trattati. I metodi, generalmente adottati per tale scopo, si basano sull'uso di informazioni ausiliarie note sui rispondenti e i non rispondenti e non tengono conto delle cause che determinano la mancata risposta. In questo lavoro, metodi che considerano tale aspetto sono stati sperimentati per il trattamento della mancata risposta dell'indagine Istat sulla Disabilità. I metodi usati sono il metodo di aggiustamento sequenziale dei pesi campionari e il metodo basato su un modello di selezione multipla del campione; il primo è stato adottato per correggere i pesi campionari dei rispondenti, il secondo per verificare le ipotesi sottostanti il primo metodo e per analizzare l'impatto degli effetti distorsivi di diverse cause di mancata risposta su alcune stime dell'indagine.

Parole chiave: Mancata risposta totale, metodo sequenziale, modello di selezione.

Abstract

The negative effects of non-response on the estimates of a sample survey must be properly treated. The methods, generally used for this purpose, are based on the use of auxiliary information known both for respondent and not respondent units, without taking into account the causes of the non-response. In this paper, methods, which consider this aspect, have been tested for the treatment of non-response in the Istat survey on Disabilities. The methods applied are the sequential weight adjustment method and the method based on a sample selection model with multiple selection equations; the first was adopted to correct the sample weights of the respondent units to the survey, the second to verify the assumptions which underlie of the first method and to analyze the impact of the bias effects produced by different causes of non-response on some survey estimates.

Keywords: total non-response, sequential weight adjustment, sample selection model.

[1] Gli autori ringraziano la Prof.ssa Daniela Cocchi per il contributo dato alla realizzazione di una parte sostanziale del lavoro (vedi De Vitiis *et al*, 2012) e i colleghi ISTAT che si sono occupati dell'indagine sulla disabilità per la loro collaborazione.

[2] Ricercatore (Istat), e-mail: devitiis@istat.it.

[3] Collaboratore di ricerca (Istat), e-mail: guandalini@istat.it.

[4] Ricercatore (Istat), e-mail: fringles@istat.it.

[5] Collaboratore di ricerca (Istat), e-mail: terribili@istat.it.

1. Introduzione

La presenza della mancata risposta totale (MRT) nelle indagini statistiche comporta una riduzione dell'attendibilità delle stime finali, determinata sia dall'aumento della varianza campionaria sia dall'introduzione di effetti distorsivi. Quest'ultimi sono tanto più gravi quanto più i rispondenti differiscono sistematicamente dai non rispondenti, rispetto a certe caratteristiche di interesse. Per eliminare, o almeno attenuare, tali effetti è necessario che, nella fase di stima di un'indagine campionaria, la MRT sia opportunamente trattata.

La mancata risposta totale può essere determinata da molteplici cause: l'irreperibilità, o mancato contatto, dovuta al fatto che l'unità statistica non ha ricevuto il modello di rilevazione o non è stata contattata dall'intervistatore; il rifiuto, quando l'unità statistica ha espressamente manifestato la volontà di non collaborare all'indagine; l'inabilità a rispondere dell'unità statistica.

Tradizionalmente, nelle indagini Istat, la MRT è trattata senza tener conto delle cause che possono generarla. Negli anni più recenti, l'attenzione dei ricercatori in ambito statistico è stata sempre più rivolta allo sviluppo di metodologie che considerano tale circostanza. L'esigenza di trattare il problema secondo un'ottica alternativa a quella tradizionale nasce da alcune importanti considerazioni: la prima è che le cause determinanti la mancata risposta hanno origine da condizioni diverse, infatti, se il rifiuto a partecipare all'indagine esplicitamente espresso da un individuo è riconducibile ad un "atteggiamento mentale", lo stesso non si può dire per il mancato contatto o per altre cause oggettive; la seconda è che se distinte cause di mancata risposta totale hanno differenti relazioni con le variabili d'indagine allora gli effetti distorsivi sulle stime possono, a loro volta, essere diversi (Groves e Couper, 1998).

In questo lavoro si propone uno studio empirico di metodi di correzione per mancata risposta totale che prendono in considerazione le diverse forme attraverso cui il fenomeno si presenta. Tali metodi partono dal presupposto che la risposta può essere vista come il risultato di distinti processi, ognuno generato da una specifica causa.

L'indagine Istat sull'"Integrazione sociale delle persone con disabilità" (indagine sulla Disabilità) del 2010 ha rappresentato il giusto contesto entro cui effettuare tale sperimentazione grazie sia alla disponibilità di informazioni ausiliarie sia alla particolare configurazione della MRT all'indagine.

La prima circostanza deriva dal fatto che l'indagine è condotta su un sotto-campione dell'indagine multiscopo "Condizioni di salute e ricorso ai servizi sanitari" anni 2004-2005 (indagine sulla Salute). Inoltre, l'indagine - realizzata con tecnica di rilevazione CATI - è affetta da un elevato tasso di mancata risposta totale, imputabile soprattutto all'irreperibilità degli individui disabili piuttosto che al rifiuto di collaborare all'indagine espresso dagli individui contattati. L'elevata quota di irreperibili è determinata dalla combinazione di più fattori: in primis, all'indagine sulla Salute non sempre erano state fornite, da parte dell'intervistato, le coordinate telefoniche, oppure quelle fornite erano errate; in secondo luogo, al momento della rilevazione, alcuni individui disabili non sono risultati raggiungibili al numero telefonico rilasciato perché cambiato o dismesso. Quest'ultima situazione è, da una parte, strettamente legata all'indagine sulla Disabilità e più precisamente al lag temporale che intercorre tra la stessa e l'indagine di riferimento sulla Salute, dall'altra, è connessa ad un aspetto critico che riguarda più in generale le indagini realizzate con tecniche CATI. Infatti, la maggior parte delle indagini basate su interviste

telefoniche, soffre da alcuni anni di un calo partecipativo legato all'aumento della sotto-copertura della rete di telefonia fissa - determinata dallo sviluppo di mezzi di comunicazione alternativi al classico telefono fisso di famiglia - soprattutto per determinate fasce di popolazione, compromettendo così la rappresentatività del campione rispetto all'intera popolazione.

I metodi presi in considerazione nella sperimentazione per il trattamento della MRT dell'indagine sulla Disabilità sfruttano informazioni ausiliarie note per i rispondenti e i non rispondenti e utilizzano, secondo un'impostazione alternativa che tiene conto delle cause che l'hanno generata, metodi generalmente adottati anche nell'approccio tradizionale al trattamento del problema.

Tali metodi sono il *metodo di aggiustamento sequenziale dei pesi campionari* (sequential weight adjustment method) e il metodo basato sul *modello di selezione multipla del campione* (sample selection model with multiple selection equations). Il primo è un metodo di aggiustamento dei pesi campionari sviluppato in più passi che utilizza tecniche di riponderazione (Rizzo *et al*., 1996; Kalton e Flores-Cervantes, 2003) basate sul "response propensity method" (Rosenbaum e Rubin, 1983; Bethlehem *et al*., 2011) o su "algoritmi di classificazione ad albero di tipo CART" (Breiman *et al*., 1984; Rizzo *et al*., 1996). Il secondo metodo costituisce uno strumento utile alla modellizzazione di differenti meccanismi di autoselezione del campione. Il modello di selezione del campione (Heckman, 1976, 1979), espresso nella forma estesa a più equazioni di selezione, assume una particolare configurazione che permette di correggere la stima di una variabile di interesse dagli effetti distorsivi generati da più cause di MRT (Groves e Couper, 1998; Bethlehem *et al*., 2011).

Entrambi i metodi considerano la natura sequenziale del processo di risposta e la distorsione come funzione di distinti processi, ma assumono ipotesi diverse circa la relazione esistente tra le fasi del processo di risposta; il metodo di aggiustamento sequenziale dei pesi campionari assume che i processi di risposta siano indipendenti, condizionatamente ad un insieme di variabili ausiliarie, mentre il metodo basato sul modello di selezione multipla del campione assume che siano correlati.

Lo studio empirico dei metodi utilizzati per la correzione della mancata risposta è stato condotto ponendoli sempre a confronto con gli stessi metodi sviluppati secondo l'impostazione tradizionale del trattamento del problema. Questo ha permesso di valutare, nel complesso, le performance delle nuove procedure rispetto a quelle standard.

Nella sperimentazione, inoltre, i due metodi sono stati utilizzati con finalità diverse; il metodo di aggiustamento sequenziale per correggere i pesi campionari dei rispondenti all'indagine sulla Disabilità; il metodo basato sul modello di selezione multipla del campione per verificare le ipotesi che sono alla base del primo metodo (indipendenza dei processi di risposta) e per analizzare l'impatto degli effetti distorsivi provocati da diverse cause di mancata risposta sulle stime di specifiche variabili dell'indagine.

L'articolo è strutturato nel modo seguente: la sezione 2 illustra le caratteristiche dell'indagine sulla Disabilità e la particolare configurazione che la MRT in essa assume; nella sezione 3 si illustrano caratteristiche e differenze dei metodi proposti, si formalizza il metodo basato sul modello di selezione multipla del campione e si descrivono i metodi di stima dei parametri del modello; le sezioni 4 e 5 riportano i risultati della sperimentazione e alcune considerazioni conclusive.

2. L'indagine "Integrazione sociale delle persone con disabilità"

2.1 Caratteristiche generali

L'indagine sull'"Integrazione sociale delle persone con disabilità" rientra nel progetto "Sistema di Informazione Statistica sulla Disabilità" nato da una convenzione tra l'Istituto nazionale di statistica e il Ministero del Lavoro e delle Politiche Sociali.

Il progetto è volto alla realizzazione di un sistema di indicatori che permette, attingendo alle diverse fonti di dati istituzionali attualmente disponibili, di monitorare il fenomeno della disabilità in Italia e di fornire un supporto alla programmazione delle politiche sociali. L'obiettivo più importante dell'indagine è di sopperire alle lacune che le altre fonti presentano sull'argomento attraverso l'acquisizione di informazioni riguardanti l'integrazione sociale delle persone con disabilità nel loro contesto di vita e le cause che ne ostacolano la piena partecipazione.

La definizione di disabile adottata nell'indagine è conforme con la nuova Classificazione internazionale del Funzionamento, della Disabilità e della Salute (Icf) approvata dall'Oms (Organizzazione Mondiale della Sanità) nel 2001. Sulla base di tale classificazione è definito disabile "chi ha una riduzione o perdita di capacità funzionale nel condurre un'attività in maniera o nei limiti considerati normali per un essere umano".

L'indagine sulla Disabilità presenta alcune importanti peculiarità: è condotta su un sotto-campione dell'indagine multiscopo "Condizioni di salute e ricorso ai servizi sanitari"[6] costituito da individui identificati, in occasione dell'indagine suddetta, come disabili; si tratta di un tipo di indagine di ritorno in quanto sono intervistate persone già contattate nell'indagine di riferimento[7]; è realizzata a distanza di 5-6 anni da quella di riferimento con la tecnica di rilevazione CATI. L'intervista è somministrata a un familiare o altro soggetto che si prende cura della persona con disabilità (proxy) in tutti i casi nei quali il disabile non è in grado di rispondere all'intervista e per i bambini disabili di età inferiore ai 14 anni.

L'indagine sulla Disabilità[8] è stata condotta per la prima volta nel 2004, l'ultima edizione risale invece al 2010.

Relativamente all'anno 2010, in conformità allo scopo individuato nel progetto suddetto, l'indagine ha acquisito numerose informazioni atte alla descrizione delle condizioni di salute e dei livelli di inclusione sociale degli intervistati nei diversi ambiti di vita (scuola, lavoro, rete di relazioni sociali, tempo libero, ecc.) e alla valutazione dell'interazione tra condizioni di salute e fattori ambientali, che possono agire come barriere (limitazioni alla mobilità, difficoltà di accesso a percorsi formativi o lavorativi, mancanza di adeguati sostegni per i bisogni assistenziali, ecc.).

La popolazione di interesse dell'indagine è stata, nell'anno 2010, diversamente definita rispetto alla prima edizione: essa è costituita dagli individui, di età compresa tra 6 e 80 anni, che all'indagine "Condizioni di Salute e ricorso ai servizi sanitari", condotta nel biennio 2004-2005, avevano riferito la propria condizione di disabilità o di avere difficoltà nelle

[6] Il disegno di campionamento è a più stadi comuni-famiglie, con stratificazione dei comuni

[7] Dall'indagine sono escluse le persone la cui disabilità è insorta successivamente al periodo di rilevazione dell'indagine salute

[8] I risultati sono presentati nelle "Statistiche in breve" del 2005

funzioni di mobilità o una riduzione di autonomia[9], ancora in quella condizione al momento dell'intervista. Sulla base della gravità delle limitazioni riferite dagli intervistati la popolazione di interesse dell'indagine risulta suddivisa in due sotto-insiemi, le persone con limitazioni funzionali gravi e le persone con limitazioni funzionali lievi.

Il campione complessivo è risultato pari a 3.502 individui con limitazioni funzionali gravi e a 7.482 individui con limitazioni funzionali lievi.

La numerosità del campione originario si è ridotta nel corso della rilevazione, perché alcuni individui identificati come disabili all'indagine sulla Salute, non sono più risultati tali: si tratta di individui usciti dalla condizione di disabilità, o deceduti oppure istituzionalizzati, ossia trasferiti in centri di ricovero in maniera stabile.

Gli individui non eleggibili hanno rappresentato il 21,6% del campione originario dei disabili con limitazioni funzionali gravi e il 15,9 del campione originario dei disabili con limitazioni funzionali lievi, comportando così una riduzione del campione complessivo che è passato da 3.502 a 2.744 unità nel primo campione, e da 7.482 a 6.293 unità nel secondo.

2.2 La mancata risposta totale

L'indagine è stata caratterizzata da un tasso di mancata risposta totale alquanto elevato, determinato in larga misura dall'irreperibilità degli individui disabili piuttosto che dal rifiuto di collaborare all'indagine espresso dagli individui contattati.

Di seguito sono riportati, con riferimento al campione dei disabili con limitazioni funzionali gravi, i risultati dell'analisi condotta sulla variabile dell'indagine "esito" utilizzata per quantificare le componenti di mancata risposta connesse a due distinte fasi del processo di risposta, la fase di contatto e la fase di partecipazione degli individui contattati.

Dalla tavola 1 risulta evidente un tasso di mancato contatto (47%) più elevato del tasso di rifiuto (23,4%) delle unità contattate. La mancata risposta ha coinvolto 1.630 unità su 2.744, di questi 1.290 non hanno potuto rispondere al follow-up dell'indagine perché è stato impossibile ricontattarli, mentre 340 si sono rifiutati di rispondere.

[9] Il collettivo contattato per l'indagine rivolta alle persone con disabilità è stato individuato tra coloro che, in occasione dell'indagine "Condizioni di salute e ricorso ai servizi sanitari" realizzata nel 2005- 2006, avevano dichiarato di:

- avere, anche con l'aiuto di ausili e apparecchi sanitari, il massimo grado di difficoltà o molta difficoltà in almeno una delle funzioni della mobilità e della locomozione (difficoltà che nelle situazioni più gravi si configura come confinamento), delle funzioni della comunicazione (vedere, sentire, parlare), delle funzioni della vita quotidiana (vale a dire delle attività di cura della persona) – rilevato per la popolazione di 6 anni e più.
- essere invalidi, secondo quanto dichiarato dagli stessi intervistati collocandosi tra i tipi di invalidità indicati (cecità, sordomutismo, sordità, invalidità da insufficienza mentale, invalidità motoria), indipendentemente dal riconoscimento legale dell'invalidità;
- avere una riduzione di autonomia, vale a dire essere colpito da una malattia cronica o da un'invalidità permanente che riduce l'autonomia personale fino ad avere bisogno di un aiuto saltuario o continuativo per le esigenze della vita quotidiana in casa o fuori casa (Istat, 2005).

Tavola 1 - Tipologia di risposta nelle due fasi del processo di risposta (contatto e partecipazione)

Fase	Esito	Numero di casi	Tasso
Prima (contatto)	Unità non contattate	1290	47,0%
	Unità contattate	1454	53,0%
	Campione effettivo	*2744*	*100,0%*
Seconda (partecipazione)	Unità partecipanti	1114	76,6%
	Unità che rifiutano	340	23,4%
	Unità contattate	*1454*	*100,0%*

Fonte: Indagine sulla Disabilità

3. Metodi per il trattamento delle componenti di mancata risposta totale

3.1 Premessa

Nel trattamento della mancata risposta totale (MRT) secondo un approccio che tiene conto delle cause che la determinano, i metodi a cui si è fatto riferimento nella sperimentazione sono il *metodo di aggiustamento sequenziale dei pesi campionari* (sequential weight adjustment method) e il metodo basato sul *modello di selezione multipla del campione* (sample selection model with multiple selection equations). Il primo metodo è stato sviluppato sia in un'ottica parametrica (Bethlehem *et al*., 2011) che non parametrica (De Vitiis *et al.*, 2012).

I metodi assumono che il processo di risposta si sviluppa in modo sequenziale attraverso un susseguirsi di fasi disposte in una struttura gerarchica e che la distorsione è funzione di contraddistinti processi generati da diverse cause di MRT.

A parte questi tratti comuni, i metodi presentano importanti differenze: assumono ipotesi diverse circa la relazione esistente tra le fasi del processo di risposta; correggono in modo differente le stime dei parametri di popolazione dagli effetti distorsivi.

Relativamente a quest'ultimo aspetto, il primo metodo porta alla costruzione di tanti fattori correttivi quante sono le fasi del processo di risposta attraverso l'uso di modelli annidati, il secondo porta alla correzione della stima di una generica variabile d'indagine attraverso l'uso di un modello che mette in relazione la variabile stessa con le fasi del processo di risposta.

Nel contesto studiato, in cui il processo di risposta è composto dalla fase di contatto delle unità campionarie e dalla fase di partecipazione all'indagine da parte delle unità contattate, il *sequential weight adjustment method* si configura come un metodo in cui la correzione dei pesi campionari è realizzata in due passi e il *sample selection model* come un modello con due equazioni di selezione.

Il metodo di aggiustamento sequenziale dei pesi campionari utilizza modelli annidati per stimare le propensioni individuali dei singoli processi e assume che le fasi del processo di risposta sono indipendenti condizionatamente ad un insieme di variabili ausiliarie (ipotesi MAR - missing at random).

Nell'approccio parametrico, il response propensity method (Rosenbaum e Rubin, 1983) è adattato alle due fasi del processo di risposta (contatto, partecipazione) attraverso l'uso di modelli annidati di tipo logit. Il modello logit, definito nella prima fase, stima le

propensioni individuali al contatto per tutte le unità del campione selezionato (modello di contatto), mentre il modello logit, definito nella seconda fase, stima le propensioni individuali alla partecipazione per le unità campionarie contattate (modello di partecipazione). Le probabilità individuali predette per le singole fasi possono essere utilizzate per la costruzione dei fattori di aggiustamento sia in modo diretto che indiretto: nel primo caso i fattori correttivi sono calcolati come inverso delle probabilità predette (response propensity weighting) con il modello di contatto (prima fase) e con il modello di partecipazione (seconda fase); nel secondo caso le probabilità predette sono utilizzate per la definizione di strati o celle di aggiustamento (response propensity stratification). Nelle celle, i fattori correttivi sono calcolati come inverso del tasso di contatto nella prima fase e del tasso di partecipazione nella seconda fase (Bethlehem et al., 2011; Groves e Couper, 1998; Iannacchione, 2003).

Nell'approccio non parametrico, l'aggiustamento sequenziale dei pesi campionari è realizzato tramite modelli basati su algoritmi di classificazione ad albero di tipo CART (Breiman *et al*., 1984; Rizzo *et al*., 1996). I modelli di classificazione sono definiti per ogni fase del processo di risposta, analogamente all'approccio parametrico. I fattori correttivi sono calcolati come inverso dei tassi di contatto e di partecipazione stimati nei nodi terminali (celle di aggiustamento) degli alberi ottimali ottenuti rispettivamente tramite il modello di classificazione del contatto e il modello di classificazione della partecipazione degli individui contattati (De Vitiis *et al.*, 2012).

L'impostazione sequenziale di aggiustamento dei pesi campionari determina, dunque, la costruzione di due fattori correttivi: il primo corregge il peso degli individui contattati per tener conto degli individui non contattati; il secondo corregge il peso dei rispondenti per tener conto dei non rispondenti tra i contattati. In questo modo i fattori correttivi catturano ognuno l'effetto distorsivo proprio associato alla singola fase del processo di risposta.

Il metodo basato sul modello di selezione del campione (Heckman, 1976, 1979) con equazioni multiple, utilizza un modello di riferimento (*equazione di regressione* o *di outcome*) per modellare i processi di selezione del campione (*equazioni di selezione*) determinati dalle fasi del processo di risposta e la media condizionata degli errori nei campioni selezionati. Il sistema di equazioni del modello è definito sul campione completo, mentre l'osservazione delle variabili dell'indagine è determinata dall'esito positivo dei processi di risposta. Il modello stima il valore atteso di una generica variabile di indagine, condizionato ad un set di variabili ausiliarie e al risultato dei processi di risposta; la sostituzione di tale valore a quello osservato della stessa variabile porta ad una stima del parametro di popolazione corrispondente corretta dagli effetti di selezione del campione (Groves e Couper, 1998; Bethlehem *et al*., 2011).

Gli effetti distorsivi determinati dal mancato contatto e dal rifiuto (modello con due equazioni di selezione) sono controllati dalle propensioni alla selezione associate alla due fasi del processo di risposta se sussiste indipendenza tra il termine di errore e le covariate del modello di regressione. I due effetti sono catturati, tramite il modello di regressione, dalla stima di specifici parametri, che sono i coefficienti delle variabili di selezione generate dalle propensioni nei singoli processi (Groves e Couper, 1998; Bethlehem *et al*., 2011).

A differenza del metodo di aggiustamento sequenziale dei pesi campionari, quest'ultimo, assume l'esistenza di correlazione sia tra i processi di risposta, sia tra questi e la variabile di indagine. Altre differenze tra i due metodi sono riconducibili all'assunzione

delle ipotesi sulle distribuzioni dei termini di errore dei modelli.

Nell'approccio sequenziale basato su modelli logit annidati, la funzione di distribuzione logistica non include l'esistenza di correlazione tra i termini di errore dei modelli, determinando così il fatto che le equazioni specificate, per ogni fase del processo di risposta, sono tra loro indipendenti e quindi stimabili separatamente. Nel modello di selezione doppia del campione i termini di errore seguono una distribuzione congiunta proprio per tener conto della correlazione tra le equazioni di selezione e tra queste e l'equazione di outcome. Per tale ragione le equazioni del sistema sono stimate simultaneamente.

La stima dei parametri dei modelli è effettuata, sia nei modelli logit annidati che nel modello di selezione, con il metodo della massima verosimiglianza (MLE). Per il modello di selezione del campione è possibile utilizzare il metodo di stima in due step proposto da Heckman (1979) che evita alcune complicazioni del metodo di stima basato sulla massima verosimiglianza completa. Tale metodo parte dalla considerazione che il valore atteso condizionato del termine di errore dell'equazione di outcome può essere visto come una variabile omessa, la cui omissione determina proprio la distorsione (Heckman, 1976, 1979).

Nel sample selection model, in particolare, la dipendenza dei metodi di stima parametrici dall'assunzione di normalità degli errori ne costituisce certamente un limite. Per superare le assunzioni sottostanti i modelli, una soluzione percorribile è quella di utilizzare metodi di stima non parametrici o semi-parametrici. Tali metodi di stima non sono stati considerati in questo lavoro, pertanto il modello è nel seguito presentato e sviluppato solo in un'ottica parametrica.

Per una descrizione più approfondita e per la formalizzazione del metodo di aggiustamento sequenziale dei pesi campionari si rinvia all'articolo di De Vitiis *et al.* (2012) oltre a quelli riportati in bibliografia.

3.2 Il metodo basato sul sample selection model

3.2.1 Formalizzazione del modello

Il *sample selection model*, introdotto da Heckman in ambito econometrico, costituisce un valido strumento per modellizzare i meccanismi di autoselezione dei rispondenti quando la mancata risposta è generata da più cause (Bethlehem *et al*., 2011).

Nel caso in cui l'osservazione di una generica variabile di indagine (variabile di outcome) dipende da diverse componenti di mancata risposta totale, il modello di selezione del campione deve tenere conto sia della natura sequenziale del processo di risposta che dei singoli processi. Nel modello devono essere definite tante equazioni di selezione quanti sono i processi di risposta coinvolti.

Se la mancata risposta totale è generata dal mancato contatto delle unità campionarie e dal rifiuto delle unità contattate allora si definiscono due equazioni di selezione su tutte le unità del campione *s*; la prima è definita per la variabile latente "propensione al contatto" (prima fase del processo di risposta) e la seconda è definita per la variabile latente "propensione alla partecipazione" (seconda fase del processo di risposta). L'equazione di outcome del modello è anch'essa definita su tutte le unità del campione *s*, ma è valorizzata soltanto quando i risultati dei processi di risposta sono congiuntamente positivi.

Nel sistema di equazioni del modello, le probabilità individuali dei due processi di selezione sono condizionate ad un insieme di variabili ausiliarie che sono inserite anche

nell'equazione di outcome. Questo perché la distorsione può essere determinata dal fatto che le propensioni alla selezione dei vari processi dipendono da variabili ausiliarie che influenzano anche la variabile d'indagine stessa.

Il modello di selezione che tiene conto della doppia selezione del campione generata dalle due fasi del processo di risposta è, dunque, specificato in termini di variabili latenti. Esso assume pertanto la seguente forma:

$$\begin{aligned} \delta_i^* &= \mathbf{X}_i^C \beta^C + \varepsilon_i^C , \\ \varphi_i^* &= \mathbf{X}_i^P \beta^P + \varepsilon_i^P , \\ y_i^* &= \mathbf{X}_i^Y \beta^Y + \varepsilon_i^Y , \end{aligned} \quad (3.1)$$

per i=1,…..,n. Le equazioni definite per le variabili latenti δ_i^* (propensione al contatto) e φ_i^* (propensione alla partecipazione) costituiscono le equazioni di selezione del modello. Se $\delta_i^* > 0$ la variabile indicatrice C_i per la i-ma unità del campione s assume valore 1, nel caso contrario assume valore 0. Se $\varphi_i^* > 0$ la variabile indicatrice P_i per la i-ma unità del campione s assume valore 1, nel caso contrario assume valore 0. La variabile indicatrice P_i è osservata soltanto quando $C_i = 1$, altrimenti è censurata. L'equazione definita per la variabile y_i^* è detta equazione di regressione (o equazione di outcome) del modello, si tratta di una variabile latente osservata soltanto quando $C_i = 1$ e $P_i = 1$. Pertanto la variabile target dell'indagine per la i-ma unità, y_i, è definita come segue

$$y_i = \begin{cases} y_i^* & se\ \ C_i = 1;\ P_i = 1 \\ . \quad se\ \ C_i = 1;\ P_i = 0\ \ oppure\ \ C_i = 0 \end{cases} \quad (3.2)$$

Le variabili esplicative del modello sono rappresentate dai vettori $\mathbf{X}_i^C$, $\mathbf{X}_i^P$ e $\mathbf{X}_i^Y$, mentre β^C, β^P e β^Y sono i coefficienti ignoti del modello. I termini di errore del modello $\left(\varepsilon_i^C, \varepsilon_i^P, \varepsilon_i^Y\right)$ sono assunti seguire una distribuzione normale multivariata $N \sim \left(\mathbf{0}, \boldsymbol{\Sigma}\right)$,

$$\begin{pmatrix} \boldsymbol{\varepsilon}_i^C \\ \boldsymbol{\varepsilon}_i^P \\ \boldsymbol{\varepsilon}_i^Y \end{pmatrix} \approx N\left(\mathbf{0}, \begin{bmatrix} 1 & \varsigma_{CP}\sigma_Y & \varsigma_{CY}\sigma_Y \\ \varsigma_{PC}\sigma_Y & 1 & \varsigma_{PY}\sigma_Y \\ \varsigma_{YC}\sigma_Y & \varsigma_{YP}\sigma_Y & \sigma_Y^2 \end{bmatrix}\right). \quad (3.3)$$

Nella matrice di varianze e covarianze, $\boldsymbol{\Sigma}$, $\varsigma_{YC}\sigma_Y$ è la covarianza tra la variabile di indagine y e il contatto, $\varsigma_{YP}\sigma_Y$ è la covarianza tra la variabile di indagine y e la partecipazione. Le correlazioni tra la variabile di indagine y e le variabili indicatrici del contatto e della partecipazione all'indagine è rispettivamente indicata con ς_{YC} e ς_{YP}, mentre la correlazione tra i tipi di risposta è indicata con $\varsigma_{PC} = \varsigma_{CP}$, essendo la matrice $\boldsymbol{\Sigma}$ simmetrica. Per tale proprietà della matrice anche $\varsigma_{YC} = \varsigma_{CY}$ e $\varsigma_{YP} = \varsigma_{PY}$.

La distorsione dovuta alla doppia selezione del campione, prima delle unità contattate e

non contattate e poi delle unità rispondenti e non rispondenti, è determinata dalla correlazione tra i termini di errore del modello, ovvero se $E\left[\varepsilon_i^Y \middle| \varepsilon_i^C\right] \neq 0$ e $E\left[\varepsilon_i^Y \middle| \varepsilon_i^P\right] \neq 0$ (Bethlehem *et al.*, 2011).

L'obiettivo del sample selection model - con due equazioni di selezione - è di stimare, tramite il modello di regressione, il valore atteso di y_i condizionato ad un set di variabili ausiliarie $\mathbf{X}_i^Y$ e al risultato dei due processi di risposta. Lo stimatore di Horvitz-Thompson per la media della popolazione di una generica variabile di indagine y, che assume la forma

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i\in s}\frac{E\left[y_i \middle| C_i = 1; P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y\right]}{\pi_i}, \quad (3.4)$$

risulta modificato in quanto il valore osservato y_i della variabile y per la *i*-ma unità del campione *s* è sostituito con la stima del valore atteso condizionato, ottenuta mediante diverse procedure di stima del modello di selezione (3.1).

La specificazione del modello per la stima dell'equazione di outcome dipende dalla natura della variabile target. Se la variabile di outcome è di tipo continuo il modello di riferimento è il modello di regressione lineare, se invece è di tipo dicotomico o categorico allora il modello di riferimento è il modello probit. Il sistema di equazioni è, in quest'ultimi casi, definito come probit multivariato con selezione del campione se la variabile di outcome è dicotomica, e come probit multinomiale con selezione del campione se la variabile di outcome è categorica.

3.2.2 Metodi di stima

Per la stima di un modello con due equazioni di selezione definite per i processi di contatto e partecipazione, è possibile utilizzare il modello proposto da Poirier (1980). In tale modello, detto "Bivariate probit model with partial observability", le variabili binarie definite per le due equazioni di selezione non sono osservate individualmente, ma ciò che è osservato è il loro prodotto. In tale ottica, il verificarsi dell'evento contatto $C_i = 1$ e dell'evento partecipazione $P_i = 1$ può essere espresso con il prodotto $C_i \times P_i = 1$. Le unità del campione *s* assumono dunque due soli valori, 1 quando $C_i = 1$ e $P_i = 1$ e 0 in tutti gli altri casi.

Il valore atteso y_i della variabile y associato alla *i*-ma unità del campione *s*, può essere espresso come

$$E\left[y_i \middle| C_i \times P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y\right] = \mathbf{X}_i^Y \beta^Y + \sigma_Y E\left[\varepsilon^Y \middle| C_i \times P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y\right], \quad (3.5)$$

dove

$$E\left[\varepsilon^Y \middle| C_i \times P_i = 1, \mathbf{X}_i^C, \mathbf{X}_i^P, \mathbf{X}_i^Y\right] = \varsigma_{YC}\sigma_Y\left(\frac{\phi\left(\mathbf{X}_i^C\beta^C\right)\Phi\left(\mathbf{X}_i^P\left(\beta^P - \varsigma_{PC}\beta^C\right)\right)\Big/\sqrt{1-\varsigma_{PC}^2}}{\Phi_2\left(\mathbf{X}_i^C\beta^C, \mathbf{X}_i^P\beta^P, \varsigma_{PC}\right)}\right) + \quad (3.6)$$

$$+\varsigma_{YP}\sigma_Y\left(\frac{\phi\left(\mathbf{X}_i^P\beta^P\right)\Phi\left(\mathbf{X}_i^C\left(\beta^C-\varsigma_{PC}\beta^P\right)\right)\Big/\sqrt{1-\varsigma_{PC}^2}}{\Phi_2\left(\mathbf{X}_i^C\beta^C,\mathbf{X}_i^P\beta^P,\varsigma_{PC}\right)}\right),$$

in cui $\phi(\cdot)$ e $\Phi(\cdot)$ rappresentano rispettivamente la funzione di densità e la cumulata della distribuzione di una normale, mentre Φ_2 è la cumulata della distribuzione della normale bivariata.

Il modello (3.5) può essere implementato stimando le due equazioni di selezione con un modello probit bivariato, oppure, ipotizzando una correlazione nulla tra i termini di errore delle due equazioni di selezione, con due modelli probit separati. In quest'ultimo caso la (3.6) può essere espressa nella forma ridotta

$$E\left[\varepsilon^Y\middle|C_i\times P_i=1,\mathbf{X}_i^C,\mathbf{X}_i^P,\mathbf{X}_i^Y\right]=\varsigma_{YC}\sigma_Y\left(\frac{\phi\left(\mathbf{X}_i^C\beta^C\right)}{\Phi\left(\mathbf{X}_i^C\beta^C\right)}\right)+\varsigma_{YP}\sigma_Y\left(\frac{\phi\left(\mathbf{X}_i^P\beta^P\right)}{\Phi\left(\mathbf{X}_i^P\beta^P\right)}\right)= \quad (3.7)$$
$$=\beta^{\lambda^C}\lambda_i^C+\beta^{\lambda^P}\lambda_i^P\ ,$$

dove $\beta^{\lambda^C}=\varsigma_{YC}\sigma_Y$ e $\beta^{\lambda^P}=\varsigma_{YP}\sigma_Y$ sono i coefficienti rispettivamente dei nuovi predittori $\lambda_i^C=\frac{\phi\left(\mathbf{X}_i^C\beta^C\right)}{\Phi\left(\mathbf{X}_i^C\beta^C\right)}$ e $\lambda_i^P=\frac{\phi\left(\mathbf{X}_i^P\beta^P\right)}{\Phi\left(\mathbf{X}_i^P\beta^P\right)}$ del modello di regressione.

Seguendo la prima procedura, si ottengono le stime $\hat{\beta}^C$ e $\hat{\beta}^P$ dei coefficienti necessari per determinare i due termini dell'equazione del valore atteso condizionato di ε_i^Y (3.6); seguendo la seconda procedura e utilizzando i parametri stimati tramite i due modelli probit per le equazioni di selezione, è possibile, invece, ottenere la stima dei termini della (3.7), $\hat{\lambda}_i^C$ e $\hat{\lambda}_i^P$, detti inverse Mills ratios (Hechman, 1979; Bethlehem *et al*., 2011).

Tali termini sono funzioni decrescenti monotone, $\hat{\lambda}_i^C$ della probabilità della *i*-ma unità del campione *s* di essere contattata e $\hat{\lambda}_i^P$ della probabilità della *i*-ma unità del campione *s* di partecipare all'indagine. Essi esprimono il fatto che le unità del campione con una elevata propensione al contatto o alla partecipazione all'indagine hanno una bassa probabilità di introdurre effetti distorsivi.

Le due procedure conducono alla stima di due covariate, i due termini dell'equazione del valore atteso condizionato di ε_i^Y nella prima e gli inverse Mills ratios nella seconda, attraverso i quali viene ridefinita l'equazione di regressione. L'introduzione nel modello di tali covariate - le variabili di selezione generate dalle propensioni nei singoli processi - consente di correggere la stima del valore atteso y_i dalla distorsione indotta dai due effetti di selezione.

4. La correzione della mancata risposta totale nell'indagine sulla Disabilità

4.1 La sperimentazione

L'indagine sulla Disabilità è affetta, come detto, da un elevato tasso di non risposta imputabile soprattutto all'elevato tasso di mancato contatto degli individui risultati disabili all'indagine sulla Salute.

Il trattamento della MRT dell'indagine in fase di stima è stato preceduto da un'analisi condotta su tre collettivi di interesse costituiti da individui non contattati, individui contattati non rispondenti e individui rispondenti.

La particolare configurazione della mancata risposta totale all'indagine e le differenze dei collettivi per alcune caratteristiche, come l'età, la ripartizione geografica di appartenenza, ecc., messe in evidenza dall'analisi riportata in De Vitiis *et al.* (2012) sono stati elementi determinanti la scelta di utilizzare un nuovo approccio per il trattamento della MRT dell'indagine. A questi si aggiunge il fatto che, grazie alle numerose informazioni rilevate all'indagine sulla Salute relativa al biennio 2004/2005, è stato possibile utilizzare, nei modelli, che sono alla base delle procedure di correzione della MRT, numerose variabili ausiliarie di tipo socio-demografico (sesso, età, stato civile, titolo di studio), oltre a quelle relative alle patologie e alle condizioni percepite dall'individuo circa le sue difficoltà nella vita quotidiana.

La sperimentazione, svolta in due fasi successive, è stata sviluppata sempre secondo due impostazioni, quella tradizionale, o standard[10], che considera come non rispondenti sia gli individui risultati irreperibili sia quelli che hanno espresso un rifiuto esplicito di collaborazione all'indagine e quella alternativa, in cui le due componenti di mancata risposta sono tenute distinte.

Nella prima fase sono state implementate diverse tecniche di riponderazione basate sia sull'uso di modelli logit che di modelli di classificazione CART. Le probabilità individuali predette tramite i modelli logit (modello di risposta nell'approccio tradizionale e modelli di contatto e di partecipazione nell'approccio alternativo) sono state utilizzate per il calcolo dei fattori correttivi applicando sia il response propensity weighting che il response propensity stratification (cfr. par. 3.1). In quest'ultimo caso, gli strati, o celle di aggiustamento, sono stati definiti tramite la tecnica degli uguali quantili delle probabilità individuali predette. Negli strati così definiti, i fattori correttivi sono stati calcolati come inverso dei tassi stimati (di risposta, contatto e partecipazione). Gli stessi tassi (De Vitiis *et al.*, 2012) sono stati calcolati nei nodi terminali (strati) degli alberi ottimali di classificazione stimati tramite i modelli CART (modello di classificazione della risposta nell'approccio tradizionale e modelli di classificazione del contatto e della partecipazione nell'approccio alternativo).

I risultati delle procedure sviluppate secondo le due impostazioni sono stati valutati attraverso un'analisi comparativa avente l'obiettivo di individuare il set di pesi finali con

[10] Relativamente ai metodi di correzione dei pesi campionari sono stati applicati modelli di stima della propensione alla risposta (modello di risposta) al fine di determinare un unico fattore correttivo. Relativamente al metodo basato sul modello di selezione del campione, nel modello è stata utilizzata un'unica equazione di selezione per la propensione alla risposta.

migliori performance. A tal fine sono stati considerati due indicatori: l'indice di concordanza e la statistica $1+CV^2$ di Kish (1992). Il primo indice, dato dalla differenza relativa tra le probabilità individuali osservate e quelle predette (De Vitiis *et al.*, 2012), è un indicatore indiretto della correzione della distorsione indotta dalla mancata risposta in quanto misura la bontà di adattamento del metodo di stima delle probabilità adottato (si fa riferimento alle probabilità individuali predette con i modelli logit e alle probabilità stimate nelle celle di ponderazione determinate sia con l'approccio parametrico che con l'approccio non parametrico). La statistica $1+CV^2$ di Kish (1992) è, invece, una misura dell'impatto della maggiore variabilità dei pesi campionari corretti per mancata risposta sulla varianza delle stime.

Nella seconda fase, per alcune variabili di interesse dell'indagine sulla Disabilità, sono stati implementati i modelli di selezione del campione, seguendo sempre l'impostazione tradizionale e alternativa; nel primo caso è stato utilizzato un modello con una sola equazione di selezione del campione definita per la propensione alla risposta, mentre nel secondo caso sono stati utilizzati modelli con due equazioni di selezione definite per la propensione al contatto e la propensione alla partecipazione (cfr. par. 3.2).

Il modello con due equazioni di selezione del campione è stato impiegato per verificare le ipotesi che sono alla base del metodo di aggiustamento sequenziale dei pesi campionari (ipotesi MAR, indipendenza dei processi di risposta). A tal fine sono state analizzate le correlazioni esistenti tra i processi di risposta e tra questi e le variabili di interesse dell'indagine.

L'analisi delle correlazioni tra le variabili di interesse e il singolo processo di risposta (modello di selezione con una equazione di selezione) o due distinti processi di risposta (modello con due equazioni di selezione) ha permesso di studiare l'impatto degli effetti distorsivi provocati da diverse cause di mancata risposta sulle stime.

Infine, per le stesse variabili, sono stati posti a confronto i valori ottenuti per le rispettive stime trattando la mancata risposta totale con tecniche di riponderazione, applicate sia in modo tradizionale che sequenziale (De Vitiis *et al*, 2012), e con il sample selection model espresso nella forma standard, ovvero con una equazione di selezione, e nella forma estesa a più equazioni di selezione.

4.2 Aggiustamento dei pesi campionari: metodo standard vs. metodo sequenziale

4.2.1 Modelli parametrici e non parametrici

I modelli logit e CART utilizzati nella sperimentazione per la costruzione dei fattori correttivi dei pesi campionari delle unità rispondenti all'indagine sono stati definiti secondo le due impostazioni. Nell'approccio tradizionale il modello di stima delle probabilità individuali è definito per la variabile risposta R_i (1.114 individui disabili rispondenti all'indagine e 1.630 individui disabili non rispondenti), mentre, nell'approccio sequenziale, i modelli di stima delle probabilità individuali di contatto (prima fase) e di partecipazione (seconda fase) sono definiti rispettivamente per le variabili contatto, C_i (1.454 individui disabili contattati e 1.290 individui disabili non contattati), e partecipazione all'indagine, P_i (1.114 individui disabili rispondenti, 340 individui disabili non rispondenti tra i contattati).

Le variabili ausiliarie utilizzate sono: presenza del telefono, età, stato civile (coniugato e non coniugato), livello di disabilità (da 1 a 3), difficoltà motorie (1=si, 0=no), numero di invalidità (da 0 a 5), numero di disabilità (da 0 a 5), difficoltà nelle funzioni giornaliere (1=si, 0=no).

Nei modelli sono state adottate diverse classificazioni della variabile età: nel modello di risposta (approccio tradizionale) sono state individuate quattro classi di età (≤12, 13-21, 22-75, >75) sia nel caso del logit che del CART; nei modelli logit di contatto e di partecipazione (approccio sequenziale) sono state individuate rispettivamente due classi di età (≤12, >12) e cinque classi di età (≤21; 22-55; 56-59; 60-77; >77), mentre, per il solo modello di classificazione (CART) della partecipazione sono state individuate tre classi di età (≤21, 22-59, >59).

Le classi di età sono state determinate tramite l'algoritmo di classificazione CART condizionando la distribuzione di ogni singola variabile target (risposta, contatto e partecipazione) ad un unico predittore costituito dalla variabile continua età.

Nella tabella che segue sono descritti i modelli adottati per la stima delle probabilità individuali in ogni approccio; in particolare, per ogni modello sono riportate le covariate risultate significative, l'AIC (*Akaike Information Criterion*) che è un indicatore di bontà di adattamento del modello logit ai dati e la funzione di costo-complessità del modello CART che costituisce un criterio di scelta ottimale dell'albero di classificazione ((Breiman *et al*., 1984; Rizzo *et al*., 1996; De Vitiis *et al*., 2012). Tali indicatori assumono valori più bassi nel modello di contatto (approccio sequenziale) rispetto al modello di risposta (approccio tradizionale).

Tavola 2 – Modelli logit e CART per la variabile risposta, contatto e partecipazione

Modello	Approccio tradizionale		Approccio sequenziale			
	Risposta		Contatto		Partecipazione	
	Covariate	Indice	Covariate	Indice	Covariate	Indice
Logit AIC	Presenza del telefono 4 classi di età Stato civile Livello di disabilità Difficoltà motorie Numero di invalidità	3.388	Presenza del telefono 2 classi di età Stato civile Difficoltà motorie Numero di invalidità Numero di disabilità	3.347	5 classi di età	1.564
CART $K_a(T)$	Presenza del telefono 4 classi di età Difficoltà nelle funzioni giornaliere	0.406	Presenza del telefono	0.325	3 classi di età	0.249

4.2.2 Principali risultati

Le tabelle che seguono mostrano alcuni importanti risultati della sperimentazione. La tavola 3 riporta i valori dell'indice di concordanza calcolato sia con riferimento ai due approcci che ai diversi metodi di stima delle probabilità individuali adottati; l'indice assume valori più elevati quando è calcolato sulle differenze tra le probabilità individuali osservate e le probabilità stimate nelle celle di ponderazione (costruite secondo la tecnica riportata in tabella) ottenute a partire dalle probabilità individuali predette dai modelli di contatto e di partecipazione utilizzati nell'approccio sequenziale.

Tavola 3 – Indici di concordanza per i modelli considerati

			Indice di concordanza		
		Tecnica	Approccio tradizionale	Approccio sequenziale	
Modello	Metodo		Risposta	Contatto	Partecipazione
Logit	Response propensity stratification	Quartili	0,569	0,574	0,645
		Quintili	0,569	0,581	
		Decili	0,573	0,584	
	Response propensity weighting	Probabilità individuali	0,565	0,569	0,647
Cart		Nodi terminali	0,574	0,583	0,648

Nelle tavole 4 e 5 sono riportate alcune informazioni di sintesi delle distribuzioni dei pesi finali, e la statistica $1+CV^2$, ottenute sempre secondo i due approcci di correzione della mancata risposta totale.

Dalla tabella 5, in cui si riportano i risultati della prima e della seconda fase di correzione nell'approccio sequenziale, si evince che la variabilità dei pesi campionari corretti nella prima fase del processo di risposta per il mancato contatto rimane sempre più contenuta rispetto a quanto accade quando si adotta un solo fattore correttivo nell'approccio tradizionale (Tav. 4).

Tavola 4 – Sintesi delle distribuzioni dei pesi finali – Approccio tradizionale

			Approccio tradizionale			
Modello	Metodo	Tecnica	Media	Max	Min	$1+CV^2$
Logit	Response propensity stratification	Quartili	1046,72	7692,57	98,83	1,680
		Quintili	1037,98	8861,92	99,02	1,673
		Decili	1037,62	9781,18	89,22	1,731
	Response propensity weighting	Probabilità individuali	1022,55	7235,38	94,09	1,615
Cart		Nodi terminali	1035,76	6796,77	94,09	**1.567**

Infine, aggiungendo un ulteriore fattore correttivo (seconda fase del processo di risposta), che tiene conto della mancata partecipazione all'indagine delle unità contattate, si nota una generale diminuzione della variabilità dei pesi finali.

E' da precisare che per la seconda fase di correzione basata sul modello logit di partecipazione, la definizione delle celle di aggiustamento (response propensity stratification) è stata effettuata considerando i soli quintili della distribuzione delle probabilità individuali predette.

Il confronto dei risultati, ottenuti con i due approcci e con una modellizzazione della risposta (o delle sue componenti) basata sia su metodi parametrici che non parametrici, mette in luce come l'approccio sequenziale conduca sempre a risultati migliori (in termini di variabilità dei pesi finali corretti), in particolare quando la tecnica di correzione è basata sugli alberi di classificazione poiché si registra una minor variabilità dei pesi corretti per le due componenti di mancata risposta.

Tavola 5 – Sintesi delle distribuzioni dei pesi finali - Approccio sequenziale

Modello	Metodo	Tecnica	Media	Max	Min	$1+CV^2$
		Prima fase				
Logit	Response propensity stratification	Quartili	800,38	5056,95	63,28	1,583
		Quintili	799,40	5597,08	61,52	1,623
		Decili	799,68	5968,48	57,55	1,664
	Response propensity weighting	Probabilità individuale	793,09	6009,83	58,63	1,603
Cart		Nodi terminali	798,18	5585,59	68,51	1,554
		Seconda fase				
Logit	Response propensity stratification	Quintili	1028,87	7081,31	104,13	1,555
	Response propensity weighting	Probabilità individuale	1027,73	7350,38	101,51	1,555
Cart		Nodi terminali	1026,71	7003,45	102,98	**1,531**

4.3 Modello di selezione del campione: modello standard vs. modello con selezione multipla

4.3.1 Modelli di selezione per specifiche variabili di interesse

Nella sperimentazione sono state individuate due variabili di tipo dicotomico su cui si è deciso di testare il metodo per ottenere stime di totali della popolazione corrette per la distorsione indotta da una o due componenti di mancata risposta totale.

Le variabili considerate riguardano, la prima, la condizione di analfabetismo degli individui disabili intervistati (Y_1=1 individui disabili analfabeti, Y_1=0 individui disabili non analfabeti) e la seconda la condizione occupazionale degli stessi (Y_2=1 individui disabili in cerca di occupazione, Y_2=0 individui disabili non in cerca di occupazione).

Nell'approccio tradizionale l'equazione di selezione, espressa nel modello per la variabile "propensione alla risposta", è descritta con riferimento alla variabile risposta R_i (1.114 unità individui disabili rispondenti e 1.630 individui disabili non rispondenti).

Nell'approccio alternativo, in cui si considerano due distinti processi di selezione, le equazioni di selezione espresse nel modello sono definite, la prima, per la variabile "propensione al contatto" e, la seconda, per la variabile "propensione alla partecipazione". Tali equazioni sono descritte con riferimento alle due variabili osservate, contatto C_i (1.454 individui disabili contattati e 1.290 individui disabili non contattati) e partecipazione all'indagine P_i (1.114 individui disabili rispondenti e 1.630 individui disabili non rispondenti). In questo caso le variabili indicatrici del contatto e della partecipazione sono definite sempre per tutte le unità del campione (2.744).

Le variabili ausiliarie utilizzate nei modelli, sia di selezione che di regressione, sono: la ripartizione geografica (Nord-Ovest, Nord-Est, Centro, Sud, Isole); difficoltà motoria (1=si, 0=no); numero di invalidità (da 0 a 5); gravità dell'invalidità (in una scala da 1 a 3); presenza del telefono (1=si, 0=no); classe d'età (fino a 12 anni, da 12 a 21 anni, da 21 a 75 anni, oltre i 75 anni).

I modelli, sia per le equazioni di selezione che per l'equazione di outcome, sono stati costruiti attraverso un'attenta scelta di queste variabili. Le variabili per ciascun modello di selezione, sono riportate nella tavola 6, in ordine di significatività.

I modelli implementati sono risultati sempre significativi, infatti il test del rapporto della massima verosimiglianza fornisce risultati positivi sulla significatività di ciascun modello. Anche i coefficienti di regressione risultano essere significativamente diversi da 0 per le variabili e per le modalità delle variabili considerate.

Tavola 6 – Modelli di selezione del campione con una equazione di selezione (risposta) e con due equazioni di selezione (contatto e partecipazione)

		Variabili di stima	
	Variabile indicatrice	Condizione di analfabetismo	Condizione occupazionale
		Covariate	
Approccio tradizionale con una equazione di selezione (unica componente di non risposta)			
Equazione di outcome		Classi d'età Numero di invalidità Ripartizione geografica Gravità dell'invalidità Difficoltà motoria	Gravità dell'invalidità Ripartizione geografica Numero di invalidità Classi d'età Difficoltà motoria
Equazione di selezione	Risposta	Presenza del telefono Classi d'età Numero di invalidità Difficoltà motoria Gravità dell'invalidità	Presenza del telefono Classi d'età Numero di invalidità Difficoltà motoria Gravità dell'invalidità
Approccio alternativo con due equazioni di selezione (due componente di non risposta)			
Equazione di outcome		Classi d'età Numero di invalidità Ripartizione geografica Gravità dell'invalidità Difficoltà motoria	Gravità dell'invalidità Ripartizione geografica Numero di invalidità Classi d'età Difficoltà motoria
1-Equazione di selezione	Contatto	Presenza del telefono Classi d'età Difficoltà motoria	Presenza del telefono Classi d'età Difficoltà motoria
2-Equazione di selezione	Partecipazione	Numero di invalidità Classi d'età Gravità dell'invalidità Ripartizione geografica	Numero di invalidità Classi d'età Gravità dell'invalidità Ripartizione geografica

5. Confronto tra metodi

Nelle tabelle 7 e 8 sono riportati i totali stimati con i diversi approcci per le modalità delle due variabili sopra descritte. Nell'approccio tradizionale i totali sono ottenuti con pesi campionari corretti con la procedura basata sul modello CART; nell'approccio sequenziale i totali sono ottenuti con pesi campionari corretti in entrambe le fasi del processo di risposta con la procedura basata sul modello CART.

Nel caso del sample selection model, i valori attesi delle variabili considerate sono determinati utilizzando diversi modelli di stima: il modello probit senza effetti indotti dalla selezione del campione, i cui parametri sono stati stimati con il metodo della massima verosimiglianza (MLE); il modello con una equazione di selezione (risposta) i cui parametri sono stati stimati con il metodo MLE e il metodo in two-step di Heckman; il modello con due

equazioni di selezione (contatto e partecipazione), dove, per la stima dei parametri delle equazioni di selezione sono state utilizzate due procedure di stima, nella prima (procedura 1) detti parametri sono stati stimati con un probit bivariato e nella seconda (procedura 2) con due probit separati (cfr. par. 3.2). Essendo le variabili dipendenti studiate di tipo dicotomico, per la stima dei parametri dell'equazione di outcome sono stati utilizzati modelli di tipo probit.

Tavola 7 – Confronto della stima del numero di individui disabili analfabeti e del numero di individui disabili non analfabeti (Valori assoluti e percentuali) nei diversi approcci

	Metodo di stima	Individui disabili analfabeti	Individui disabili non analfabeti	Totale
Approccio basato sulla correzione dei pesi campionari				
Tradizionale (una fase)	CART	97.088 (8,1%)	1.104.100 (91,9%)	1.201.188 (100.0%)
Sequenziale (due fasi)	CART	96.034 (8,0%)	1.105.154 (92,0%)	1.201.188 (100.0%)
Approccio basato sul modello probit senza selezione				
	MLE	97.836 (7,9%)	1.102.881 (92,1%)	1.200.717 (100.0%)
Approccio basato sul sample selection model				
1 eq. selezione	MLE	93.709 (7,4%)	1.107.479 (92,6%)	1.200.716 (100.0%)
	Heckman two-step	89.348 (7,6%)	1.111.369 (92,4%)	1.200.717 (100.0%)
2 eq. selezione	Procedura 1	91.003 (7,6%)	1.109.714 (92,4%)	1.200.717 (100.0%)
	Procedura 2	91.049 (7,6%)	1.109.627 (92,4%)	1.200.717 (100.0%)

Tavola 8 – Confronto della stima del numero di individui disabili in cerca di occupazione e del numero di individui disabili non in cerca di occupazione (Valori assoluti e percentuali) nei diversi approcci

	Metodo di stima	Individui disabili in cerca di occupazione	Individui disabili non in cerca di occupazione i	Totale
Approccio basato sulla correzione dei pesi campionari				
Tradizionale (una fase)	CART	326.178 (27,1%)	875.010 (72,9%)	1.201.188 (100.0%)
Sequenziale (due fasi)	CART	322.670 (26,9%)	878.518 (73,1%)	1.201.188 (100.0%)
Approccio basato sul modello probit senza selezione				
	MLE	323.078 (27,9%)	877.639 (72.1%)	1.200.717 (100.0%)
Approccio basato sul sample selection model				
1 eq. selezione	MLE	327.850 (27,3%)	873.339 (72,7%)	1.201.189 (100.0%)
	Heckman two-step	329.825 (27,5%)	870.892 (72,5%)	1.200.717 (100.0%)
2 eq. selezione	Procedura 1	330.640 (27,5%)	870.077 (72,5%)	1.200.717 (100.0%)
	Procedura 2	329.776 (27,5%)	870.941 (72,5%)	1.200.717 (100.0%)

Per le due variabili considerate nella sperimentazione si verificano situazioni opposte. Rispetto al modello probit senza effetti di selezione, i metodi che ne tengono conto portano a valori più bassi nel caso della stima del numero di disabili analfabeti, e a valori più alti nel caso della stima del numero di individui disabili in cerca di occupazione.

Il motivo di questo si può apprezzare maggiormente quando si utilizza il sample selection model: nel caso in cui si considera una sola equazione di selezione la correlazione tra la variabile di interesse e la mancata risposta ς_{YR}, che dà il segno alla correzione della distorsione, nel primo caso è negativa, ς_{Y_1R} = - 0.0250, mentre nel secondo caso è positiva, ς_{Y_2R} = 0.0137.

Il discorso è analogo quando si considerano due equazioni di selezione. In questo caso entrambe le componenti di mancata risposta hanno una correlazione negativa con la variabile "condizione di analfabetismo", infatti ς_{Y_1C} = -0.0253 e ς_{Y_1P} = -0,0005, e positiva con la variabile "condizione occupazionale", ς_{Y_2C} = 0,0151 ς_{Y_2P} = 0.0008. Il segno assunto dalle correlazioni tra la variabile di interesse e i processi di selezione determina il segno della correzione nella (3.6) e (3.7).

Analizzando le stime ottenute con i due modelli di selezione è possibile affermare che l'effetto di selezione, dovuto ai processi di selezione del contatto e della partecipazione, ha un impatto sulle stime della percentuale di individui disabili analfabeti e degli individui disabili in cerca di occupazione rispettivamente vicino allo 0,5% ed allo 0,3%.

L'ordine di grandezza della correzione di tali effetti distorsivi è influenzato dal livello di correlazione tra la mancata risposta o le sue componenti con la variabile di interesse. Nell'esempio queste, seppur basse, hanno un effetto non trascurabile.

Inoltre, poiché i valori delle correlazioni ς_{YR} e ς_{YC}, per entrambe le variabili indagate nei due modelli di selezione sono molto simili, è possibile dedurre che gli effetti distorsivi della mancata risposta totale siano in gran parte determinati dalla componente mancato contatto e, solo per una parte residuale, dalla componente rifiuto. Quindi possiamo affermare che l'effetto di selezione è principalmente dovuto al mancato contatto e non al rifiuto a partecipare all'indagine.

La correlazione positiva tra la le due componenti di mancata risposta denota una buona propensione alla partecipazione all'indagine degli individui disabili, una volta contattati. Tuttavia, essendo questa esigua, la correzione che ne deriva ha un impatto marginale sulle stime. Questo risultato, determinato molto probabilmente dall'elevato tasso di mancato contatto, giustifica l'ipotesi di indipendenza tra le due componenti di mancata risposta fatta nel lavoro di De Vitiis *et al*. (2012).

6. Conclusioni

La sperimentazione di metodi alternativi al trattamento della MRT è stata resa possibile dalla disponibilità di un ampio numero di variabili ausiliarie note per le unità rispondenti e le unità non rispondenti. La carenza di informazione ausiliaria può costituire, in generale, un limite applicativo dei metodi presentati nel lavoro che, tuttavia, in futuro, potrà essere superato grazie alla crescente disponibilità di sistemi integrati di informazioni di fonte amministrativa che costituiscono un punto centrale della modernizzazione avviata dall'Istituto.

L'applicazione dell'approccio sequenziale di aggiustamento dei pesi campionari ha dato buoni risultati, soprattutto quando i fattori correttivi sono stati determinati a partire da una modellizzazione non parametrica dei processi di risposta.

La sperimentazione del metodo di correzione della mancata risposta totale basato sul *sample selection model* ha consentito di analizzare gli effetti distorsivi determinati dai legami tra le diverse componenti della mancata risposta e tra queste e le specifiche variabili di interesse.

Il *sample selection model*, soprattutto nella sua formulazione estesa a più equazioni di selezione, è un approccio molto interessante perché applicabile a diversi contesti di studio, come la stima per indagini basate su tecniche miste di rilevazione. Molte indagini Istat, infatti, stanno introducendo tale metodologia che, se da un lato è utilizzata proprio per contenere la mancata risposta totale, dall'altro può introdurre specifici effetti distorsivi sulle stime che devono essere analizzati e trattati in fase di stima. Sebbene il metodo presenti livelli di complessità elevati, esso consente di studiare gli effetti combinati della tecnica di rilevazione e delle mancata risposta totale, o delle sue componenti.

L'applicazione del *sample selection model* al contesto presentato nel lavoro costituisce una fase iniziale di studio che ci ha consentito di intuire le potenzialità del metodo. Ulteriori approfondimenti in un'ottica simulativa, tuttavia, sono necessari per poter studiare le proprietà degli stimatori utilizzati. Inoltre, sarà opportuno valutare anche il ricorso a metodi di stima non parametrici e semi-parametrici che possono portare a notevoli vantaggi nel caso in cui non sono verificate le ipotesi alla base dei modelli.

Infine, è nei nostri obiettivi l'applicazione del modello di selezione multipla del campione a situazioni più complesse in cui più fattori possono concorrere a introdurre effetti distorsivi sulle stime delle indagini statistiche (mixed-mode, MRT o sue componenti).

Riferimenti bibliografici

Bethlehem, J., Cobben, F. e Schouten, B. 2011. *Handbook of Nonresponse in household surveys*. New York: Wiley.

Breiman, L., Friedman, J.H., Olshen, R.A. e Stone, C.J. 1984. *Classification Regression Trees*. Belmont: Wadsworth International Group.

De Vitiis, C., Cocchi, D., Inglese, F. e Terribili M.D. 2012. *Treatment of total nonresponse via sequential weight adjustment in the Italian disability survey*. Italian Journal of Applied Statistics, Special Issue, Vol. 24, N. 1.

Groves, R.M. e Couper, M.P. 1998. *Nonresponse in household interview surveys*. New York: Wiley.

Heckman J. J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement 5, 475–492*.

Heckman, J. J. 1979. Sample selection bias as specification error. *Econometrica*, 47: 153-161.

Iannacchione, V.G. 2003. Sequential weight adjustments for location and cooperation propensity for 1995 national survey of family growth. *Journal of Official Statistics*, 19: 31-43.

ISTAT 2012. Inclusione sociale delle persone con limitazioni dell'autonomia personale. Statistiche report. http://www.istat.it/it/archivio/77546.

Kalton, G. e Flores-Cervantes, I. 2003. Weighting methods. *Journal of Official Statistics*, 19: 81-97.

Kish, L. 1992. Weighting for Unequal Pi. *Journal of Official Statistics*, 8: 183-200.

Poirier, D.J. 1980. Partial observability in bivariate probit models. *Journal of Econometrics*, 12: 210-217.

Rosenbaum, P.R. e Rubin, D.B. 1984. Reducing the bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association*, 79, 516-524.

Rizzo, L., Kalton, G. e Brick, J.M. 1996. A comparison of some weighting adjustment methods for panel nonresponse, *Survey Methodology*, 22: 43-53.