istat
working
papers

N. 3
2015

Il metodo sequenziale di aggiustamento delle componenti della mancata risposta totale sperimentato nell'indagine Istat sulla disabilità

Claudia De Vitiis, Francesca Inglese, Marco D. Terribili

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 3
2015

Il metodo sequenziale di aggiustamento delle componenti della mancata risposta totale sperimentato nell'indagine Istat sulla disabilità

Claudia De Vitiis, Francesca Inglese, Marco D. Terribili

Istat

Istat Working Papers

Il metodo sequenziale di aggiustamento delle componenti della mancata risposta totale sperimentato nell'indagine Istat sulla disabilità

N. 3/2015

ISBN 978-88-458-1830-1

Il metodo sequenziale di aggiustamento delle componenti della mancata risposta totale sperimentato nell'indagine Istat sulla disabilità[1]

Claudia De Vitiis, Francesca Inglese, Marco D. Terribili[2]

Sommario

Nella fase di stima di un'indagine campionaria, le tecniche di ponderazione adottate per compensare gli effetti negativi della mancata risposta totale sulle stime, si basano sull'uso di informazioni ausiliarie note sui rispondenti e i non rispondenti, generalmente senza distinguere i diversi tipi di mancata risposta. Il lavoro si propone di trattare separatamente le componenti della mancata risposta, in particolare il mancato contatto e il rifiuto, con l'obiettivo di ridurre gli effetti distorsivi della mancata risposta totale. Il metodo adottato, noto in letteratura come metodo di aggiustamento sequenziale, utilizza per la costruzione di fattori correttivi delle componenti di mancata risposta totale modelli annidati specificati per ogni fase del processo di risposta. L'ipotesi su cui si basa tale approccio è che i diversi tipi di risposta sono indipendenti condizionatamente ad un insieme di variabili ausiliarie.

Parole chiave: mancata risposta totale, metodo sequenziale, modelli annidati

Abstract

The weighting adjustment techniques adopted in the estimation phase of a sample survey to compensate negative total non-response effects on the estimates are based on the use of auxiliary information known both for respondents and non-respondents, usually without distinguishing among different types of non-response. This paper proposes to treat the components of non-response, i.e. impossible to contact and refusal, separately with the aim of reducing non-response bias. The adopted method, known in the literature as sequential weight adjustment, uses nested models for the construction of the adjustment factors of the components of non-response, specified for each phase of the response process. The underlying assumption is that, conditional on auxiliary information, the different types of non-response are independent.

Keywords: total non-response, sequential adjustment, nested models

[1] Una versione preliminare di questo lavoro è stata presentata alle Giornate della ricerca in Istat del 10-11 novembre 2014. Le opinioni espresse sono solo degli autori e non implicano la responsabilità dell'Istat.

[2] devitiis@istat.it, inglese@istat.it, terribili@istat.it.

Indice

1. Introduzione

L'oggetto di questo lavoro si colloca all'interno dell'importante tematica concernente lo studio di metodi statistici idonei ad eliminare, o più realisticamente attenuare, gli effetti negativi della mancata risposta totale nelle indagini statistiche.

La mancata osservazione del fenomeno indagato presso alcune unità statistiche chiamate a partecipare alla rilevazione comporta una riduzione dell'attendibilità delle stime finali, determinata sia dall'aumento della varianza di campionamento sia dall'introduzione di effetti distorsivi.

Nella fase di stima di un'indagine campionaria, per attenuare gli effetti distorsivi determinati dal fenomeno della mancata risposta totale si adotta una metodologia che consiste nella modifica dei pesi campionari associati alle unità rispondenti, affinché essi possano tener conto delle unità non rispondenti.

L'introduzione di fattori correttivi della mancata risposta totale trova fondamento teorico nel campionamento in più fasi (Särndal, 1992), dove al meccanismo probabilistico di selezione del campione di prima fase si aggiunge, nelle fasi successive, un nuovo tipo di casualità determinata dalla probabilità di risposta ignota assunta da ogni unità del campione. In sostanza, introducendo un modello di risposta casuale, la teoria di campionamento basata sul disegno diventa una teoria detta *quasi-randomization* (Oh and Sheuren, 1983).

I metodi di correzione dei pesi associati alle unità incluse nel campione dei rispondenti, detti *tecniche di riponderazione*, sono fondati sull'uso di informazioni ausiliarie (Rizzo *et al*., 1996; Kalton and Flores-Cervantes, 2003). Nel caso in cui si disponga di un insieme di variabili ausiliarie per tutte le unità campionarie, rispondenti e non rispondenti, la costruzione dei fattori correttivi può essere perseguita attraverso una modellizzazione esplicita della probabilità di risposta, assumendo l'esistenza di un legame funzionale tra la propensione delle unità del campione ad essere rispondenti e le caratteristiche da esse possedute.

Come è noto, la mancata risposta totale può essere determinata da diverse cause (irreperibilità o mancato contatto, rifiuto, ecc.) ma, generalmente, di questo non si tiene conto nell'applicazione dei metodi di aggiustamento dei pesi diretti.

In questo lavoro si propone, con riferimento all'indagine ISTAT "Integrazione sociale delle persone con disabilità" del 2010, una soluzione alternativa a quella standard di trattamento del problema in cui le componenti di mancata risposta totale sono trattate separatamente. Il metodo proposto, noto in letteratura come *sequential weight adjustment* (Groves and Couper, 1998; Bethlehem *et al.*, 2011), considera anche un particolare aspetto del processo di risposta, ovvero la sua natura sequenziale.

Secondo tale prospettiva il processo di risposta si sviluppa in più fasi e la partecipazione di un individuo all'indagine si concretizza nell'ultima fase di un processo caratterizzato da una sequenza di eventi, le diverse fasi del processo di risposta, ognuno annidato nel precedente. In tale approccio, le componenti di mancata risposta sono gerarchicamente distinte e indipendenti; i fattori correttivi sono determinati, per ogni fase del processo di risposta, sulla base di specifici modelli tra loro annidati.

L'applicazione di una metodologia innovativa non ancora sperimentata nell'Istituto per il trattamento del problema esposto costituisce il punto focale del lavoro. Tuttavia l'elemento di novità apportato all'approccio in due fasi consiste nell'utilizzo di un metodo non parametrico, basato sugli alberi di classificazione (Breiman *et al*., 1984; Rizzo *et al*., 1996), per la costruzione sequenziale dei fattori correttivi delle componenti di mancata risposta totale.

L'articolo è strutturato nel modo seguente: nella sezione 2 si discute delle cause della mancata risposta e della distorsione introdotta, nella stima dei parametri di interesse di un'indagine, dalle varie componenti; la sezione 3 illustra il metodo di aggiustamento della mancata risposta totale nell'approccio sequenziale, sia con riferimento a modelli parametrici che non parametrici; la sezione 4 presenta i risultati della sperimentazione condotta per l'indagine sulla Disabilità in un'ottica comparativa tra approccio sequenziale e approccio standard; la sezione 5 riporta alcune considerazioni conclusive e indica le possibili linee di ricerca future sul tema della mancata risposta totale.

2. Cause ed effetti della mancata risposta totale

Con il termine mancata risposta totale si indica la circostanza che un'attesa risposta, da parte dell'unità statistica chiamata a partecipare alla rilevazione, per un qualsiasi motivo, non perviene all'ente promotore della rilevazione stessa (Särndal *et al*., 2005).

La mancata risposta totale può essere determinata da molteplici cause, le principali sono: l'*irreperibilità*, o *mancato contatto*, determinata dal 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 determinata, ad esempio, da condizioni psico-fisiche.

L'identificazione di differenti tipi di mancata risposta totale è rilevante nella fase di prevenzione del fenomeno (Groves and Couper, 1998) per la predisposizione di azioni di intervento specifiche, ma altrettanto lo è nella successiva fase di stima. Gli effetti della distorsione sulle stime possono, infatti, variare a seconda delle cause che determinano la mancata risposta totale. D'altra parte 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.

Gli effetti della distorsione della mancata risposta totale sugli stimatori di parametri della popolazione possono essere analizzati attraverso due approcci basati, il primo, su un modello di risposta deterministico e, il secondo, su un modello di risposta casuale (Särndal *et al*., 1992).

Il modello di risposta deterministico assume che la popolazione è suddivisa in due strati composti rispettivamente dalle unità rispondenti e dalle unità non rispondenti. Se il processo di risposta è assunto essere di tipo sequenziale e, per semplicità, le componenti della mancata risposta totale sono costituite dal mancato contatto e dal rifiuto, allora la popolazione risulta suddivisa, nella prima fase, in unità contattate e unità non contattate e, nella seconda fase, in unità rispondenti e non rispondenti tra quelle contattate. In tale ottica, la distorsione dello stimatore va analizzata considerando le sotto-popolazioni definite per le due fasi del processo e anche la circostanza che la partecipazione, o il rifiuto, delle unità a collaborare all'indagine risulta condizionata dall'esito positivo della fase in cui le unità sono contattate.

L'impatto della distorsione sulle stime è difficile da misurare in quanto dovrebbe essere disponibile un campione osservato dalla popolazione delle unità non contattate e un campione osservato dalla popolazione delle unità non rispondenti tra quelle contattate.

Con riferimento alla popolazione di dimensione *N*, se si considera come parametro di interesse la media della popolazione, $\overline{Y}$, della variabile *y*, nel modello di risposta deterministico la distorsione della stima $\hat{\overline{Y}}_{s_R}$, ottenuta dal campione dei soli rispondenti, può essere scomposta rispetto alle componenti della mancata risposta totale:

$$B\left(\hat{\overline{Y}}_{s_R}\right)=\frac{N_{NC}}{N}\left(\frac{N_{R|NC}}{N}\left(\overline{Y}_R-\overline{Y}_{R|NC}\right)+\frac{N_{RF|NC}}{N}\left(\overline{Y}_R-\overline{Y}_{RF|NC}\right)\right)+\frac{N_C}{N}\frac{N_{RF|C}}{N_C}\left(\overline{Y}_R-\overline{Y}_{RF|C}\right), \quad (1)$$

dove: N_{NC} è il numero delle unità non contattate nella popolazione, che può essere diviso a sua volta in due parti, $N_{R|NC}$ e $N_{RF|NC}$, che rappresentano rispettivamente il numero delle unità rispondenti e il numero delle unità non rispondenti nella popolazione composta dalle unità non contattate; $\overline{Y}_{R|NC}$ e $\overline{Y}_{RF|NC}$ sono rispettivamente le medie dei due gruppi nella popolazione; N_C indica il numero delle unità contattate nella popolazione; $N_{RF|C}$ è il numero delle unità della popolazione che hanno espresso un rifiuto se contattate; $\overline{Y}_{RF|C}$ è la media della variabile target *y* sulle unità che rifiutano di partecipare all'indagine anche se contattate (Bethlehem *et al*., 2011).

L'ultimo termine dell'equazione (1) esprime il condizionamento della seconda fase del processo

di risposta (la partecipazione o il rifiuto delle unità a collaborare all'indagine) all'esito positivo della prima fase in cui le unità sono contattate.

La distorsione dello stimatore dipende, dunque, da più fattori, ovvero i tassi di mancato contatto e di rifiuto e le differenze tra le varie medie afferenti alle unità appartenenti agli strati della popolazione definiti per le due fasi del processo di risposta.

3. Il metodo di aggiustamento sequenziale

L'approccio sequenziale al trattamento della mancata risposta totale pone al centro dell'attenzione proprio le diverse cause della mancata risposta e mira alla riduzione della distorsione delle stime attraverso la costruzione di fattori correttivi definiti per ogni fase del processo di risposta.

Quando la mancata risposta è caratterizzata da due componenti, ad esempio "mancato contatto" e "rifiuto", il processo di risposta risulta suddiviso in due fasi e i fattori di aggiustamento sono definibili, il primo come un correttore del mancato contatto - attraverso il quale i pesi campionari degli individui risultati reperibili sono modificati per rappresentare gli irreperibili- e, il secondo come un correttore della mancata partecipazione all'indagine degli individui contatti - che modifica ulteriormente i pesi degli individui rispondenti per rappresentare i non rispondenti tra i reperibili.

Una tecnica di trattamento sequenziale delle componenti della mancata risposta totale consiste nell'adattare il *response propensity method* (Rosenbaum and Rubin, 1983; Bethlehem *et al.*, 2011) in modo da riflettere le diverse fasi del processo di partecipazione degli individui all'indagine. In questo caso possono essere utilizzati modelli logistici annidati (*sequential two-stage propensity model adjustments*), ovvero modelli specificati separatamente per ogni fase del processo di risposta, in cui le probabilità degli individui ad essere contatti (prima fase) e le probabilità degli individui contattati di partecipare all'indagine (seconda fase) sono condizionate ad un set di variabili ausiliarie (Bethlehem *et al.*, 2011; Groves and Couper, 1998; Iannacchione, 2003). L'assunzione fondamentale su cui è basato tale approccio è che le fasi del processo di risposta sono indipendenti condizionatamente a un insieme di variabili ausiliarie (MAR - *missing at random*).

Le probabilità individuali predette tramite i modelli definiti per le due fasi del processo di risposta, possono essere utilizzate per la costruzione dei fattori di aggiustamento direttamente o indirettamente: nel primo caso i due fattori correttivi sono calcolati come inverso delle probabilità di contatto predette nella prima fase e delle probabilità di risposta predette nella seconda fase per le unità contattate (*response propensity weighting*); nel secondo caso, le probabilità individuali predette tramite i modelli annidati, sono utilizzate per la costruzione di strati o celle di aggiustamento (*response propensity stratification*).

La costruzione sequenziale dei fattori correttivi della mancata risposta totale attraverso modelli non parametrici può essere basata sugli algoritmi di classificazione ad albero (CART). 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 stimati nei nodi terminali (celle) degli alberi di classificazione risultati ottimali nel modello di classificazione della variabile target *contatto* (prima fase) e nel modello di classificazione della variabile target *partecipazione* all'indagine delle unità contattate (seconda fase).

3.1. Costruzione sequenziale dei fattori correttivi con modelli parametrici e non parametrici

Al fine di descrivere e formalizzare l'approccio sequenziale al trattamento della mancata risposta totale, è utile introdurre la seguente notazione simbolica. Si indichi con U la popolazione oggetto di interesse e si supponga di aver selezionato, mediante un determinato disegno di campionamento, un campione s di dimensione n ($i=1,\ldots,n$) a cui è associata una misura di probabilità $p(s)$. Sia π_i la probabilità di inclusione di primo ordine relativa alla generica unità della popolazione U. Si indichi inoltre con s_C ($s_C \subseteq s$) il campione costituito dalle unità contattate nella prima fase e con s_P ($s_P \subseteq s_C$) il campione delle unità che partecipano all'indagine nella seconda

fase se contattate.

Nella prima fase, la probabilità della *i*-ma unità del campione *s* di essere contattata, condizionatamente alle caratteristiche $\mathbf{X}_i^C = (X_{i1}, X_{i2}, ..., X_{iq})'$, può essere espressa come

$$\theta_{1i} = \theta_1\left(\mathbf{X}_i^C\right) = P\left(C_i = 1 \middle| \mathbf{X}_i^C\right), \qquad (2)$$

dove C_i assume valore 1 se la *i*-ma unità del campione *s* è contattata e valore 0 nel caso contrario. Nella seconda fase, condizionatamente al risultato della fase precedente, ovvero quando C_i=1, la probabilità che la *i*-ma unità del campione s_C partecipi all'indagine, condizionatamente alle caratteristiche $\mathbf{X}_i^P = (X_{i1}, X_{i2}, ..., X_{iv})'$, è

$$\theta_{2i} = \theta_2\left(\mathbf{X}_i^P\right) = P\left(P_i = 1 \middle| \mathbf{X}_i^P, C_i = 1\right), \qquad (3)$$

dove P_i assume valore 1 se la *i*-ma unità appartenente al campione s_C partecipa all'indagine e valore 0 nel caso contrario.

3.1.1. Modelli logistici annidati

Nell'approccio sequenziale, per la costruzione di fattori correttivi, possono essere utilizzati modelli annidati di tipo logit, uno per ogni fase del processo di risposta.

Il modello logit per la probabilità di contatto (prima fase) è

$$log\left(\frac{C_i}{1-C_i}\right) = logit\left(\theta_1\left(\mathbf{X}_i{}^C\right)\right) = \mathbf{X}_i{}^C{}'\boldsymbol{\beta}^C \quad (i=1,\ldots,n) \qquad (4)$$

per ogni *i*-ma unità appartenente al campione iniziale *s* di dimensione *n*, che rappresenta il numero di unità del campione per le quali è definito il vettore $\mathbf{C} = \left(C_1, C_2, ..., C_n\right)'$. La probabilità di contatto è stimata come

$$\hat{\theta}_{1i} = \hat{\theta}_1\left(\mathbf{X}_i^C\right) = \frac{exp\left(\mathbf{X}_i^C\hat{\boldsymbol{\beta}}^C\right)}{1+exp\left(\mathbf{X}_i^C\hat{\boldsymbol{\beta}}^C\right)}. \quad (5)$$

Il modello logit per la probabilità di partecipazione all'indagine (seconda fase) è

$$log\left(\frac{P_i}{1-P_i}\right) = log\,it\left(\theta_2\left(\mathbf{X}_i^{(P|C=1)}\right)\right) = \mathbf{X}_i^{(P|C=1)}{}'\boldsymbol{\beta}^{(P|C=1)} \quad (i=1,...,n_C) \quad (6)$$

per ogni *i*-ma unità appartenente al campione s_C di dimensione n_C, che rappresenta il numero di unità contattate per le quali è definito il vettore $\mathbf{P} = \left(P_1, P_2, ..., P_{n_C}\right)'$. La probabilità di partecipazione all'indagine è stimata come

$$\hat{\theta}_{2i} = \hat{\theta}_2\left(\mathbf{X}_i^P\right) = \frac{exp\left(\mathbf{X}_i^{P|C=1}\ \hat{\boldsymbol{\beta}}^{P|C=1}\right)}{1 + exp\left(\mathbf{X}_i^{P|C=1}\ \hat{\boldsymbol{\beta}}^{P|C=1}\right)}. \quad (7)$$

L'utilizzo diretto o indiretto delle probabilità individuali predette per la costruzione dei fattori correttivi conduce a due diverse formulazioni dello stimatore del parametro di interesse: lo stimatore *response propensity weighting* e lo stimatore *response propensity stratification*. Se il parametro di interesse è la media della popolazione $\bar{Y}$ della variabile *y*,

$$\bar{Y} = \frac{1}{N}\sum_{i \in U} y_i\,, \quad (8)$$

uno stimatore diretto (HT, di Horvitz-Thompson) di $\bar{Y}$ può essere espresso, nel campionamento in più fasi, in funzione della probabilità di inclusione π_i associata alla *i*-ma unità del campione *s*, della probabilità di contatto θ_{1i} e della probabilità di partecipazione all'indagine delle unità contattate θ_{2i},

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i \in s_P} \frac{y_i}{\pi_i \theta_{1i} \theta_{2i}} \quad (i=1,\ldots,n_P). \qquad (9)$$

Lo stimatore *response propensity weighting* si ottiene sostituendo nella espressione (9) la probabilità di contatto θ_{1i} e la probabilità di partecipazione θ_{2i} rispettivamente con la propensione al contatto, $\hat{\theta}_1\left(\mathbf{X}_i^C\right)$, e la propensione alla partecipazione, $\hat{\theta}_2\left(\mathbf{X}_i^P\right)$,

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i \in s_P} \frac{y_i}{\pi_i\, \hat{\theta}_1\left(\mathbf{X}_i^C\right)\hat{\theta}_2\left(\mathbf{X}_i^P\right)} \quad (i=1,\ldots,n_P). \qquad (10)$$

I fattori di aggiustamento per la *i*-ma unità sono specificati come reciproco della propensione al contatto $\hat{\theta}_1\left(\mathbf{X}_i^C\right)$ e come reciproco della propensione alla partecipazione $\hat{\theta}_2\left(\mathbf{X}_i^P\right)$

$$\gamma_{1i} = \frac{1}{\hat{\theta}_1\left(\mathbf{X}_i^C\right)} \quad \text{e} \quad \gamma_{2i} = \frac{1}{\hat{\theta}_2\left(\mathbf{X}_i^P\right)}. \quad (11)$$

Se le probabilità individuali predette sono usate per la costruzione di celle, o strati (*response propensity stratification*) allora i due fattori correttivi del peso diretto sono ottenuti, il primo come inverso della probabilità di contatto stimata nella cella *f* (*f*=1,…, *F*) e il secondo come inverso della probabilità di partecipazione stimata nella cella *g* (*g*=1,…, *G*). Per la cella *f* definita sul campione *s* (*i*=1,…,*n*) e per la cella *g* definita sul campione s_C ($i = 1,...,n_C$) i fattori di aggiustamento sono rispettivamente

$$\gamma_{1f}^{'} = \left(\hat{\theta}_f\right)^{-1} = \left(\frac{n_{C,f}}{n_f}\right)^{-1} \quad \text{e} \quad \gamma_{2g}^{'} = \left(\hat{\theta}_g\right)^{-1} = \left(\frac{n_{P,g}}{n_{C,g}}\right)^{-1}, \qquad (12)$$

dove: la probabilità di contatto stimata nella cella f, $\hat{\theta}_f$, è l'*f*-mo elemento scalare di $\hat{\boldsymbol{\Theta}}^C = \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_F,\right)$, $n_{C,f}$ è il numero delle unità contattate nella cella (prima fase) e n_f è il numero totale di unità nella cella; la probabilità di partecipazione stimata nella cella *g*, $\hat{\theta}_g$, è il *g*-mo elemento scalare di $\hat{\boldsymbol{\Theta}}^P = \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_G,\right)$, $n_{P,g}$ è il numero delle unità rispondenti, essendo contattate, nella cella *g* (seconda fase) e $n_{C,g}$ il numero delle unità contattate nella stessa cella.

In questo caso si realizzano due post-stratificazioni: la prima assegna lo stesso fattore di aggiustamento alle unità contattate in ogni strato definito sul campione *s*; la seconda assegna lo stesso fattore di aggiustamento alle unità rispondenti in ogni strato definito sul campione s_c. Lo stimatore *response propensity stratification* calibra il campione *s* utilizzando la propensione al contatto e successivamente calibra il campione s_c utilizzando la propensione alla partecipazione delle unità contattate.

3.1.2. Modelli CART

Un modello CART descrive la distribuzione condizionata di una variabile target **Z** dato un set di *p* predittori definiti dalla matrice **X** di dimensione $n \times p$. Questo modello ha due principali componenti: un albero *T* con *M* nodi terminali, e un parametro $\boldsymbol{\Theta} = \left(\theta_1, \theta_2, ..., \theta_M\right)$ che associa il valore del parametro θ_m con l' *m*-mo nodo terminale (*m=1*,…, *M*). Un modello di decisione ad albero è completamente specificato dalla coppia $\left(T, \boldsymbol{\Theta}\right)$. Se **X** giace nella regione corrispondente all'*m*-mo nodo terminale, allora $\mathbf{Z}|\mathbf{X}$ ha distribuzione $f = \left(\mathbf{Z}|\theta_m\right)$, *f* è usata per rappresentare la distribuzione condizionata in θ_m. Il modello è detto di regressione o classificazione ad albero a seconda che **Z** sia quantitativa o qualitativa.

L'idea sottostante tale approccio è di suddividere in modo ricorsivo le unità del campione in gruppi sempre più piccoli fino ad ottenere la partizione migliore dove la variabile target raggiunge la massima omogeneità.

Generalmente, la strategia preferita per la ricerca della partizione ottimale è di considerare un albero completo *T* e di potare l'albero utilizzando una funzione detta "costo-complessità". La scelta dell'albero di classificazione ottimale è basata su un compromesso tra la complessità dell'albero e l'errore di classificazione delle unità nei gruppi e corrisponde all'albero con un valore minimo della funzione costo-complessità $\Phi_\alpha(T)$:

$$\Phi_\alpha\left(T\right) = \Phi\left(T\right) + \alpha \cdot Q\left(T\right), \qquad (13)$$

in cui $\Phi(T)$ è l'errore di classificazione associato ad un dato albero *T*, *Q(T)=M* è il numero di nodi terminali e α è un coefficiente che penalizza la complessità dell'albero. L'errore di classificazione $\Phi(T)$ assume il valore zero nella massima partizione e tende all'infinito quando l'albero coincide con il nodo radice.

Il parametro $\alpha \geq 0$ controlla il trade-off tra la dimensione dell'albero e la bontà di adattamento ai dati. A valori elevati di α corrispondono alberi di piccole dimensioni, il contrario accade per valori piccoli di α. Per $\alpha = 0$ la soluzione della funzione è l'albero pieno *T*. Per ottenere l'albero finale $T_{\hat{\alpha}}$ è necessario stimare il coefficiente α, che minimizza la funzione costo-complessità $\Phi_{\alpha}(T)$. La stima, $\hat{\alpha}$, è ottenuta attraverso ripetute analisi di cross-validation.

Nell'approccio sequenziale di aggiustamento della mancata risposta totale è necessario definire due modelli CART annidati, uno per ogni fase del processo di risposta: il primo, con parametri $\left(T^C, \boldsymbol{\Theta}^C\right)$, per la stima della probabilità delle unità del campione *s* di essere contattate e il secondo, con parametri $\left(T^P, \boldsymbol{\Theta}^P\right)$, per la stima della probabilità di partecipazione all'indagine delle unità contattate.

Nella prima fase la distribuzione condizionata di $\mathbf{C}|\mathbf{X}^C$ per l'*m*-mo nodo terminale è definita come $f_1 = \left(\mathbf{C}|\theta_m\right)$, dove **C** è il vettore delle variabili indicatrici del contatto definito sul campione *s* di dimensione *n* (*i*=1,…,*n*) e $\mathbf{X}^C$ è la matrice dei predittori di dimensione $n \times q$ (modello CART di contatto); nella seconda fase la distribuzione condizionata di $\mathbf{P}|\mathbf{X}^P$ per l'*m*-mo nodo terminale è definita come $f_2 = \left(\mathbf{P}|\theta_m\right)$, dove **P** è il vettore delle variabili indicatrici della partecipazione definito sul campione s_C di dimensione n_c (*i*=1,…,n_c) e $\mathbf{X}^P$ è la matrice dei predittori di dimensione $n_C \times v$ (modello CART di partecipazione).

I parametri stimati nei modelli di classificazione delle variabili target contatto e partecipazione sono: l'albero finale $T^C_{\hat{\alpha}}$ con *L* (*L*<*M*) nodi terminali (*l*=1,…,*L*) e il vettore dei tassi di contatto stimati, $\hat{\boldsymbol{\Theta}}^C = \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_L\right)$, nel primo modello; l'albero finale $T^P_{\hat{\alpha}}$ con *J* (*J*<*M*) nodi terminali (*j*=1,…,*J*) e il vettore dei tassi di partecipazione stimati, $\hat{\boldsymbol{\Theta}}^P = \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_J\right)$, nel secondo modello.

I fattori di aggiustamento sono calcolati per le due fasi del processo di risposta: il primo fattore correttivo è dato dall'inverso del tasso di contatto stimato in ciascun nodo terminale individuato tramite il modello di classificazione del contatto; il secondo fattore correttivo è dato dall'inverso del tasso di partecipazione stimato in ciascun nodo terminale individuato tramite il modello di classificazione della partecipazione. I due fattori di aggiustamento sono, in sostanza, ottenuti come espresso in formula (12).

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

L'indagine sull'"Integrazione sociale delle persone con disabilità" (Istat, 2012) ha l'obiettivo di acquisire informazioni riguardanti il livello di integrazione dei disabili nella società e le cause che ne ostacolano la piena partecipazione.

L'indagine presenta alcune peculiarità che hanno reso possibile la sperimentazione di un nuovo approccio metodologico al trattamento della mancata risposta.

Si tratta, infatti, di un'indagine di ritorno condotta nel 2010 sul sotto-campione degli individui risultati disabili all'indagine multiscopo "Condizioni di salute e ricorso ai servizi sanitari" (realizzata cinque anni prima). Tale circostanza rende disponibili numerose informazioni sui rispondenti e i non rispondenti, individuabili tra le variabili rilevate nella prima occasione di indagine; si tratta di informazioni relative a caratteristiche di tipo socio-demografico degli individui

o che riguardano la condizione di inabilità o disabilità degli individui.

Date le particolari caratteristiche della popolazione oggetto di osservazione, l'indagine ha previsto, per gli individui con disabilità o inabilità gravi, la possibilità che le risposte fossero fornite da familiari o da persone che se ne prendevano cura. Tale soluzione è stata adottata al fine di evitare che la mancata partecipazione degli individui disabili all'indagine fosse determinata dalle gravi condizioni degli stessi.

L'indagine sulla Disabilità è stata realizzata con una tecnica di rilevazione di tipo CATI ed è affetta da un elevato tasso di mancata risposta totale, imputabile principalmente all'irreperibilità degli individui disabili (mancanza o dismissione del telefono fisso che era stato fornito nel 2010), più che al rifiuto di collaborare all'indagine espresso dagli individui contattati.

Dalla tavola che segue, in cui sono riportati gli esiti dell'indagine nelle due fasi del processo di risposta, risulta evidente l'elevato tasso di mancato contatto (47%) rispetto al tasso di rifiuto (23,4%) delle unità contattate.

Tavola 1 – Tipologie di risposta nelle due fasi del processo

FASI	Esiti	Numero di casi	Tasso
	Unità non contattate	1290	47,0%
Prima	Unità contattate	1454	53,0%
	Campione effettivo	*2744*	*100,0%*
	Unità partecipanti	1114	76,6%
Seconda	Unità che rifiutano	340	23,4%
	Unità contattate	*1454*	*100,0%*

Fonte: Indagine sulla Disabilità

4.1. La sperimentazione - principali risultati

La sperimentazione realizzata per la correzione dei pesi base associati alle unità rispondenti (corrispondenti ai pesi finali dell'indagine del 2010) è stata sviluppata secondo due impostazioni, l'approccio di aggiustamento sequenziale e l'approccio standard in una singola fase.

I risultati delle due procedure sono stati valutati attraverso un'analisi comparativa, estesa su più livelli, con l'obiettivo di individuare il set di pesi finali con migliori performance. A tal fine sono stati considerati due indicatori: l'indice di concordanza, dato dalla differenza relativa tra le probabilità individuali osservate e quelle predette sulla base dei criteri scelti, che è un indicatore indiretto della correzione della distorsione indotta dalla mancata risposta; la statistica $1+CV^2$ di Kish (1992) che è una misura dell'impatto della maggiore variabilità dei pesi campionari corretti per mancata risposta sulla varianza delle stime.

In entrambi gli approcci, la costruzione dei fattori correttivi è stata realizzata a partire dai modelli parametrici e non parametrici utilizzati (modello di risposta nell'approccio tradizionale e modelli di contatto e di partecipazione nell'approccio sequenziale) e utilizzando metodi e tecniche diverse.

Quando sono stati utilizzati modelli di tipo logit, la costruzione dei fattori correttivi è stata realizzata seguendo i due approcci descritti nel paragrafo 3.1.1.; in particolare, la costruzione delle celle di aggiustamento (response propensity stratification) è stata effettuata con la *tecnica degli uguali quantili* delle probabilità individuali predette. Quando sono stati utilizzati, invece, modelli di tipo CART, i fattori correttivi sono stati calcolati nei nodi terminali, o celle, degli alberi di classificazione risultati ottimali, come descritto nel paragrafo 3.1.2.

Le tabelle che seguono presentano alcuni importanti risultati della sperimentazione: la tavola 2, relativa ai modelli adottati per la stima delle probabilità individuali, riporta 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. Tali indicatori assumono valori più bassi nel modello di contatto (approccio sequenziale) rispetto al modello di risposta (approccio tradizionale).

La tavola 3 presenta i valori dell'indice di concordanza calcolato per ogni approccio e criterio adottato; l'indice assume valori più elevati quando è calcolato sulle differenze tra le probabilità individuali osservate e quelle predette sulla base dei modelli di contatto e di partecipazione utilizzati nell'approccio sequenziale.

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 $\Phi_\alpha(T)$	Presenza del telefono 4 classi di età Difficoltà nelle funzioni giornaliere	0.406	Presenza del telefono	0.325	3 classi di età	0.249

Tavola 3 – Indici di concordanza per i modelli considerati

MODELLO	Metodo	Tecnica	Indice di concordanza		
			Approccio tradizionale	Approccio sequenziale	
			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 individuale	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 seguendo rispettivamente l'approccio tradizionale e l'approccio sequenziale al trattamento 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 (Tab. 4).

Inoltre, aggiungendo nella seconda fase del processo di risposta (Tab. 5) il fattore correttivo 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.

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

MODELLO	Metodo	Approccio tradizionale				
		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 individuale	1022,55	7235,38	94,09	1,615
Cart		Nodi terminali	1035,76	6796,77	94,09	**1.567**

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

MODELLO	Metodo	Prima fase				
		Tecnica	Media	Max	Min	$1+CV^2$
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**

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 che l'approccio sequenziale conduce sempre a risultati migliori (in termini di variabilità dei pesi finali corretti) e che, in particolare, la tecnica di correzione sequenziale basata sugli alberi di classificazione è, nel nostro caso, da preferire anche perché la variabilità dei pesi finali risulta complessivamente meno incrementata dall'introduzione di due fattori correttivi.

5. Conclusioni e sviluppi futuri

I risultati della sperimentazione sono molto incoraggianti, tanto da suggerire l'estensione di tale approccio, ove possibile, ad altre indagini dell'Istituto e l'approfondimento di altri metodi di correzione della mancata risposta che tengano conto delle diverse cause del fenomeno.

E' sicuramente nei nostri obiettivi, infatti, la sperimentazione del metodo *sample selection model* nella forma estesa con equazioni di selezione multiple legate alle componenti della mancata risposta totale (Bethlehem *et al*., 2011). L'aspetto interessante di questo metodo è l'assunzione dell'esistenza sia di correlazione tra i tipi di risposta, sia della relazione tra la variabile risposta e la variabile di interesse dell'indagine di cui si tiene conto nella modellizzazione, sempre restando sotto l'ipotesi di *ignorabilità* del meccanismo della mancata risposta (MAR). L'approfondimento, inoltre, vorrebbe andare anche nella direzione dello studio di metodi che consentono un superamento di tale ipotesi di *ignorabilità*.

Il *sample selection model* è, inoltre, un approccio applicabile al trattamento della mancata risposta totale nelle indagini mixed-mode quando si dispone di informazioni specifiche sulla singola modalità di rilevazione (utilizzo di *paradati*), sia nel caso in cui le diverse tecniche di rilevazione sono concomitanti che nel caso in cui sono sequenziali.

Il metodo presenta livelli di complessità elevati ma consente di incorporare nella stima di parametri della popolazione gli effetti distorsivi introdotti dalla tecnica di rilevazione e dalle componenti di mancata risposta totale.

Lo studio e lo sviluppo di nuove metodologie per il trattamento della mancata risposta totale costituisce sicuramente una sfida interessante e auspicabile, anche considerando le possibilità offerte dalla crescente disponibilità nell'Istituto di sistemi integrati di informazioni di fonte amministrativa, da una parte, e dall'aumento della complessità delle indagini, dall'altra, dovuto all'utilizzo di tecniche di rilevazione di tipo mixed-mode. Tali tecniche, se da un lato sono messe in atto proprio per contenere la mancata risposta totale, dall'altro possono introdurre degli specifici effetti distorsivi sulle stime che è opportuno analizzare e trattare.

Riferimenti bibliografici

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

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

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

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., 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.

Oh, H.L. , Scheuren, F. 1983. Weighting adjustments for unit nonresponse, in *Incomplete Data in Sample Surveys. Theory and Bibliographies*, W.G. Madow, I. Olkin & D.B. Rubin, 2:143-184, New York: Academic Press.

Rosenbaum, P.R. and 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. and Brick, J.M. 1996. A comparison of some weighting adjustment methods for panel nonresponse, *Survey Methodology*, 22: 43-53.

Särndal, C.E., Swensson, B. and Wretman, J.H. 1992. *Model Assisted Survey Sampling*, Cap XV, New York: Springer.

Särndal, C.E., Lundström, S. 2005. *Estimation in surveys with nonresponse*. New York: Wiley.