LA QUALITÀ NEI PROCESSI DI DATA CAPTURING. IL CASO DELL'INDAGINE SUGLI ASPETTI DELLA VITA QUOTIDIANA

Claudio Ceccarelli, Marco Fortini, Manuela Murgia, Alessandra Nuccitelli, Rita Ranaldi, Francesca Rossetti

1. Introduzione

La qualità dell'informazione prodotta da un'indagine statistica dipende da numerosi fattori: definizioni e concetti adottati, disegno di campionamento, tecniche di indagine, strumenti, via via fino alle modalità di analisi e rappresentazione dei dati. Il concetto di Total Survey Error (TSE) è sviluppato secondo questo principio per legare la qualità di ciascuna fase di processo a quella dell'informazione prodotta, in modo da poter monitorare e intervenire sulle attività di indagine che risultano più critiche in un'ottica di rapporto tra costi e benefici (Lyberg and Stukel, 2017). In questo approccio la qualità dell'informazione è rappresentata dalla somma degli scostamenti, in termini di distorsione e variabilità, tra il valore stimato e il valore vero, che intervengono durante le diverse fasi del processo.

Questo lavoro si concentra sul processo di raccolta dei dati, determinante nella misura in cui contenerne l'errore contribuisce in modo sostanziale a migliorare la qualità complessiva dell'informazione prodotta.

La prestazione del processo di raccolta dei dati può essere valutata tramite dati riguardanti l'esecuzione delle fasi di processo, i cosiddetti paradati (Kreuter, 2013). Prendono il nome di paradati tutte le informazioni sull'andamento del processo di produzione di un'indagine statistica raccolte nel corso del processo stesso. Il contatto, la partecipazione, il numero di solleciti, la modalità e la durata dell'intervista (o della compilazione) sono tutti esempi di paradati; ad essi si affiancano altre informazioni trattate in occasione di un'indagine e costituite da: dati; metadati; dati ausiliari; macrodati (o stime).

L'analisi dei paradati e della loro associazione con le altre informazioni disponibili si dimostra utile a costruire indicatori di prestazione delle fasi di indagine per orientare i necessari aggiustamenti di processo.

Il classico esempio di paradato è costituito dalla variabile indicatrice di risposta dell'unità statistica, a partire dalla quale si può calcolare il tasso di risposta, inteso come rapporto tra il numero di rispondenti e il totale delle unità statistiche

selezionate nel campione. Se il tasso di risposta è costante rispetto a una data variabile di interesse Y, allora la mancata risposta si definisce 'ignorabile' rispetto a Y e l'analisi riferita ai soli rispondenti risulterà non distorta. Dato che questa misura non si può ottenere in modo diretto perché Y nella pratica non è nota sui non rispondenti, si può ripiegare su un vettore X di variabili ausiliarie e/o altri paradati, noto per tutte le unità del campione, condizionatamente al quale la variabile Y sia indipendente dal tasso di risposta. In questo caso si dirà che la mancata risposta è ignorabile rispetto a Y, condizionatamente al vettore X (Little and Rubin, 2002).

Mentre l'associazione tra il vettore X e il tasso di risposta è studiata sui dati in essere, l'indipendenza condizionata tra il tasso di risposta e la variabile Y viene postulata sulla base di informazioni e conoscenze esterne. L'analisi del tasso di risposta (o propensione alla risposta) condizionatamente al profilo dei rispondenti aiuta a identificare i gruppi più a rischio di mancata risposta ed è utile sia ai fini di calibrazione delle stime, sia per interventi correttivi da effettuare in corso d'indagine, sia, infine, per la revisione del disegno di indagine nel suo complesso.

Analogamente alla propensione alla risposta è possibile considerare altri paradati, come:

- eleggibilità, che misura la quantità e la tipologia dei casi non eleggibili erroneamente inclusi nelle liste di indagine;
- tecnica di indagine (per le indagini *mixed-mode*), per studiare se la modalità di risposta influenza il risultato della misura;
- numero di solleciti, per studiare quanti farne e su quali sottogruppi;
- profili di risposta critici, per tenere sotto controllo particolari tendenze come, ad esempio, segnalare sempre la prima modalità di risposta o privilegiare traiettorie che abbreviano il tempo di compilazione del questionario.

La variabilità dei paradati può anche essere messa in relazione a gruppi significativi, come le unità statistiche assegnate ad uno stesso rilevatore o a un ufficio territoriale, per evidenziare criticità locali sulle quali è opportuno intervenire.

Parte di queste analisi mira a creare correttori delle stime tramite modelli che tengono conto delle diverse fonti di errori campionari e non campionari. Tale approccio è complicato dalla necessità di disporre di informazione ridondante, di ricorrere a ipotesi spesso difficili da verificare in pratica, considerando per giunta solo una o poche fonti di errore alla volta. In questa sede si considerano, piuttosto, le determinanti della qualità per migliorare le future occasioni di indagine o per modificare in maniera flessibile l'indagine in corso, raccogliendo all'origine dati più completi ed affidabili a parità di costo.

Come esempio dell'approccio proposto, viene descritta un'analisi della propensione a rispondere oppure a utilizzare una delle due tecniche previste (CAWI

[1] o CAPI-PAP[2]) per l'indagine sugli 'Aspetti della vita quotidiana' (AVQ) condotta nel 2019. I paradati in questione sono messi in relazione a profili significativi delle unità di rilevazione per suggerire adattamenti nella strategia di contatto e sollecito.

Nel paragrafo 2 si descrivono l'indagine e i dati utilizzati, mentre nel terzo viene illustrato il modello logistico adottato per l'analisi e si presentano i principali risultati. Infine, nel paragrafo 4 sono delineati i possibili metodi e strumenti per il miglioramento della qualità della rilevazione nell'ambito di una strategia che coinvolge i diversi attori che ruotano intorno al complesso processo di indagine.

2. Dati

L'analisi presentata nel paragrafo 3 è stata effettuata sfruttando le informazioni disponibili relativamente alla fase di raccolta dei dati dell'Indagine Multiscopo sulle Famiglie: Aspetti della vita quotidiana (AVQ), con riferimento all'anno 2019. Questa indagine rientra nel Programma statistico nazionale ed ha lo scopo di rilevare annualmente le informazioni necessarie per conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno e se sono soddisfatti del funzionamento di quei servizi di pubblica utilità che dovrebbero contribuire al miglioramento della qualità della vita.

Le informazioni vengono raccolte attraverso una tecnica mista, che si avvale di un questionario online da auto-compilare da parte dei rispondenti (tecnica CAWI) oppure di un'intervista diretta con questionario elettronico, somministrato da un intervistatore, e contestuale consegna di un questionario cartaceo da auto-compilare da parte dei componenti della famiglia (tecnica CAPI-PAP). In particolare, le due tecniche sono applicate in sequenza: per l'indagine dell'anno 2019, infatti, le famiglie hanno potuto compilare il questionario online (tecnica CAWI) nel periodo dal 28 febbraio al 31 marzo 2019, utilizzando le credenziali riportate nella lettera di invito. Qualora una famiglia non avesse avuto la possibilità di rispondere all'indagine tramite Internet, al termine del periodo previsto un intervistatore comunale si è recato presso l'abitazione della famiglia stessa (tecnica CAPI-PAP), per rivolgere le stesse domande del questionario online a tutti i suoi componenti. Sono stati fatti due solleciti postali durante il periodo previsto per la compilazione con tecnica CAWI; per le famiglie che avevano già fatto accesso al questionario web senza concludere l'intervista, i solleciti sono stati effettuati tramite e-mail.

[1] Computer-Assisted Web Interviewing.
[2] Computer-Assisted Personal Interviewing e Paper and Pencil.

Il campione teorico del 2019 è formato da 25.177 famiglie residenti in 783 comuni. In particolare, rispetto al totale delle famiglie del campione, il 35% circa ha fatto accesso al portale web e, fra queste, quasi il 90% ha portato a termine l'intervista con tecnica CAWI. Alla fine del periodo previsto per la rilevazione con tecnica CAWI, le famiglie non rispondenti – ad esclusione di quelle che, pur avendo contattato il *contact center*, sono risultate essere non eleggibili e di quelle che avevano iniziato la compilazione del questionario senza concluderla – sono state oggetto di intervista con tecnica CAPI-PAP; delle 16.348 famiglie complessive coinvolte, il 71,5% ha terminato l'intervista.

In questa prima sperimentazione sono stati utilizzati i soli dati disponibili e/o raccolti nell'ambito della fase di conduzione d'indagine: per ciascuna famiglia del campione teorico sono disponibili informazioni ausiliarie relative alla famiglia anagrafica (ad esempio, il numero di componenti, la dimensione demografica del comune di residenza della famiglia, la provincia di residenza) o relative al singolo membro della famiglia (il sesso, l'età, la cittadinanza, il paese e la provincia di nascita, il ruolo nella famiglia[3]). Le informazioni raccolte attraverso il questionario non sono al momento disponibili.

La rilevazione è stata condotta avvalendosi di due differenti sistemi di gestione di indagine, uno per la parte CAWI e uno combinato per la parte CAPI-PAP. Il sistema di gestione della tecnica CAPI-PAP prevede la raccolta delle informazioni relative ai tentativi di contatto, al numero di visite effettuate dall'intervistatore presso la famiglia ma, non essendo l'intervistatore obbligato alla registrazione, questi dati possono essere incompleti o parziali: la loro disponibilità potrebbe essere invece utile per integrare e approfondire le analisi sulla propensione alla risposta da parte delle famiglie.

È importante inoltre ricordare che circa il 97% (24.465) delle famiglie del campione di AVQ era già presente nel campione dell'Indagine da lista (L) della rilevazione censuaria 2019 sulla popolazione; le restanti famiglie sono state estratte dalle Liste anagrafiche comunali (Lac) del comune di appartenenza, in quanto il comune nel Censimento era incluso nella sola indagine areale. La presenza della quasi totalità delle famiglie nelle due indagini campionarie permetterà, in una fase successiva di analisi, l'accesso ad un maggior numero di variabili ausiliarie e quindi l'applicazione di altri strumenti di analisi e modelli della non-risposta.

[3] Intestatario della Scheda di Famiglia (ISF), coniuge dell'ISF o altro membro.

3. Analisi della propensione alla risposta

3.1. Modello utilizzato

L'analisi dei dati riguarda le 23.093 famiglie eleggibili[4] dell'indagine AVQ per l'anno 2019 per le quali si studia l'associazione tra il comportamento all'intervista – "Risponde con tecnica CAWI" (CW), "Risponde con tecnica CAPI-PAP" (CP), "Non risponde" (NR) – e alcune caratteristiche disponibili nella lista di partenza. A questo scopo viene utilizzato un modello di regressione logistica multinomiale (o politomica) ad effetti fissi.

Le variabili ausiliarie inserite nel modello sono le seguenti[5]:

- *Ripartizione geografica* ("Nord-ovest"; "Nord-est"; "Centro"; "Mezzogiorno");
- *Classe di popolazione del comune* ("<=5.000 abitanti"; "5.001-50.000 abitanti"; "50.001-150.000 abitanti"; ">=150.001 abitanti");
- *Cittadinanza dell'ISF* ("Italiana"; "Straniera")
- *Età dell'ISF* ("<50; ">=50");
- *Numero di componenti della famiglia anagrafica* ("<=2"; ">2").

Un'altra caratteristica riferita all'ISF, come il *Sesso*, non figura nel modello finale in quanto è risultata avere un impatto debolmente significativo sul fenomeno in esame.

Il modello utilizzato – privo di effetti interattivi – si articola in due equazioni, cioè quante sono le categorie di comportamento previste, al netto della categoria scelta come riferimento (CP):

$$\begin{aligned} \ln\left(\frac{\Pr(Y_i = \mathrm{CW}|\mathbf{x}_i)}{\Pr(Y_i = \mathrm{CP}|\mathbf{x}_i)}\right) &= \alpha_{\mathrm{CW}} + \beta_{\mathrm{CW}}^{11} x_i^{11} + \beta_{\mathrm{CW}}^{21} x_i^{21} + \cdots + \beta_{\mathrm{CW}}^{53} x_i^{53} \\ \ln\left(\frac{\Pr(Y_i = \mathrm{NR}|\mathbf{x}_i)}{\Pr(Y_i = \mathrm{CP}|\mathbf{x}_i)}\right) &= \alpha_{\mathrm{NR}} + \beta_{\mathrm{NR}}^{11} x_i^{11} + \beta_{\mathrm{NR}}^{21} x_i^{21} + \cdots + \beta_{\mathrm{NR}}^{53} x_i^{53} \end{aligned} \tag{1}$$

avendo indicato con $\mathbf{x}_i = (x_i^{11}, x_i^{21}, x_i^{31}, x_i^{41}, x_i^{42}, x_i^{43}, x_i^{51}, x_i^{52}, x_i^{53})$ il vettore dei regressori indicatori[6] osservati sulla famiglia i – relativi, nell'ordine, alle variabili

[4] Sono state escluse dall'analisi anche le famiglie che non è stato possibile contattare per errori di lista.

[5] La classificazione dell'*Età dell'ISF* e del *Numero dei componenti della famiglia anagrafica* è stata scelta sulle base di analisi preliminari dell'associazione con la variabile dipendente.

[6] Ogni variabile esplicativa viene qui rappresentata utilizzando tanti regressori indicatori quante sono le modalità della variabile meno una, scelta come categoria di riferimento. Ad esempio, la variabile *Ripartizione geografica*, con quattro modalità, viene rappresentata nel modello utilizzando i tre regressori indicatori seguenti: X^{51} (= 1 se Centro, 0 altrimenti), X^{52} (= 1 se Nord-est, 0 altrimenti), X^{53} (= 1 se Nord-ovest, 0 altrimenti), avendo assunto "Mezzogiorno" come categoria di riferimento.

Numero di componenti della famiglia anagrafica, *Età dell'ISF*, *Cittadinanza dell'ISF*, *Classe di popolazione del comune*, *Ripartizione geografica* – e con le lettere greche i parametri da stimare.

3.2. Risultati

Nella Tabella 1 è riportata la variazione dell'adattamento del modello ai dati ottenuta tralasciando una variabile esplicativa alla volta e mantenendo invece le altre (statistica del Chi-quadrato), con il *p-value* associato.

I risultati mostrano che ciascuna delle variabili ha un effetto significativo sulla propensione di una famiglia a un certo comportamento (colonna 3); in particolare, il livello della statistica del Chi quadrato (colonna 2) evidenzia che la *Ripartizione geografica* è il fattore più rilevante tra quelli considerati. Seguono, staccati per importanza, la *Classe di popolazione del comune* e la *Cittadinanza dell'ISF*. Infine, l'*Età dell'ISF* e il *Numero di componenti della famiglia anagrafica*, sembrano avere un impatto più limitato sul fenomeno in esame.

Tabella 1 – *Variabili esplicative per statistica del Chi-quadrato di Wald e p-value associato.*

Variabile	Chi quadrato di Wald	Pr > ChiQuad
Ripartizione geografica	1.461,81	<,0001
Classe di popolazione del comune	489,30	<,0001
Cittadinanza dell'ISF	393,45	<,0001
Età dell'ISF	47,29	<,0001
Numero di componenti della famiglia anagrafica	47,22	<,0001

Fonte: nostre elaborazioni su dati Istat

Quando tutte le variabili coinvolte nel modello sono categoriche, come nel caso in questione, l'interpretazione dei risultati può risultare più immediata esaminando direttamente gli *Odds Ratio*[7] (OR), anziché i parametri in (1).

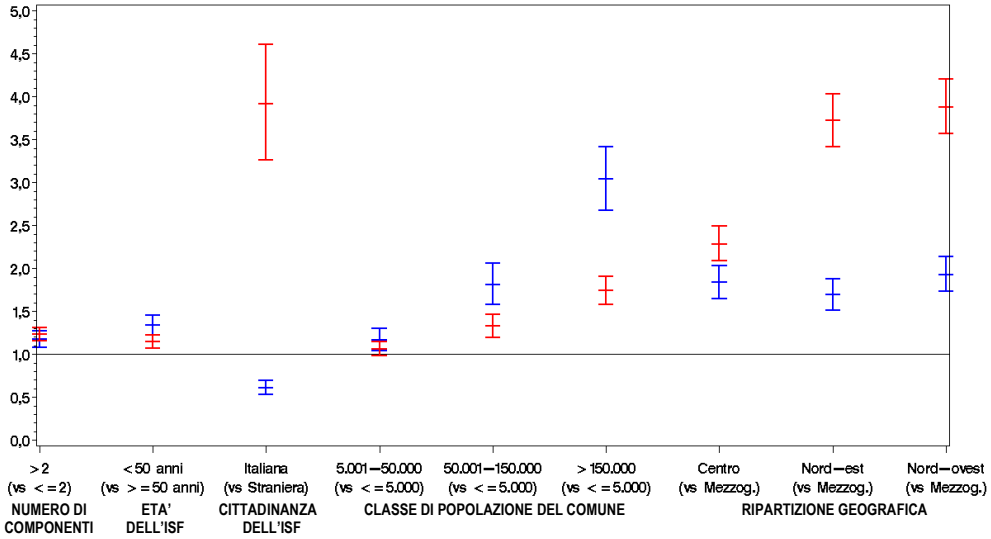
Nella Figura 1 per ciascuna variabile esplicativa sono riportate le stime di massima verosimiglianza degli OR e dei rispettivi intervalli di confidenza al 95%, relative alle due equazioni del modello. Tali stime permettono una caratterizzazione

[7] L'OR non è altro che il rapporto tra gli *Odds* relativi al confronto tra due categorie della variabile risposta – ad esempio, CW e CP – in due situazioni alternative di una delle variabili esplicative (ad esempio, "Centro" e "Mezzogiorno", se si considera la variabile *Ripartizione geografica*). In altre parole, l'OR permette di valutare immediatamente di quanto cresce o decresce il rischio che la variabile di risposta assuma una certa categoria (CW), anziché un'altra scelta come riferimento (CP), a seguito della variazione del valore assunto dalla variabile esplicativa – ad esempio, da "Mezzogiorno" a "Centro", se si sceglie "Mezzogiorno" come modalità di riferimento – e al netto degli effetti di tutte le altre variabili.

delle famiglie secondo la propensione alle varie tipologie di comportamento prese in considerazione.

Nel contesto in esame, un interesse specifico è rivestito dalla categoria CW, che gioca un ruolo particolarmente importante nel contenimento del tasso di mancata risposta e dell'errore di misurazione e, più in generale, nel miglioramento del processo di raccolta dei dati.

Figura 1 – *Odds Ratio e intervalli di confidenza al 95% per i regressori indicatori relativi al modello (1)*



Nota: *la modalità di riferimento per ogni variabile ausiliaria è indicata tra parentesi
Fonte: nostre elaborazioni su dati Istat

Specificamente, la propensione a rispondere con tecnica CAWI, piuttosto che con modalità CAPI-PAP, risulta maggiore per le famiglie:

- del Nord (OR superiori a 3,70) e, in misura più contenuta, per quelle del Centro (OR = 2,28);
- con ISF italiano (OR = 3,83), soprattutto se ha un'età inferiore ai 50 anni (OR = 1,15);
- di almeno tre componenti (OR = 1,23).

Per di più, la propensione alla tecnica CAWI tende a crescere all'aumentare della dimensione comunale (OR = 1,73 per la classe di popolazione più ampia).

Tuttavia, all'aumentare della dimensione comunale, si riscontra una tendenza ancora più marcata a non rispondere all'indagine (OR pari a 1,16 e 1,80 per le classi di popolazione intermedie; superiore a 3 per le famiglie che vivono in comuni con

più di 150.000 abitanti). Inoltre, il rischio di mancata risposta risulta più elevato per le famiglie:

- con ISF straniero (OR = 0,61), soprattutto se ha un'età inferiore ai 50 anni (OR = 1,34);
- di almeno tre componenti (OR = 1,17).

Infine, vale la pena evidenziare che le famiglie del Mezzogiorno mostrano un comportamento più incline alla risposta, con una maggiore propensione ad utilizzare la tecnica CAPI-PAP.

Alla luce dei risultati ottenuti, per la prossima edizione dell'indagine, potrebbe essere proposta una modifica dell'impianto di rilevazione, passando da un approccio puramente sequenziale del *mixed-mode* a uno concorrente in cui le tecniche sono in campo contemporaneamente e invitando al CAWI o al CAPI-PAP i profili di popolazione più propensi a ciascuna tecnica. Inoltre, per favorire l'utilizzo della modalità CAWI, si potrebbe valutare l'introduzione di una strategia di sollecito (alternativa alla lettera cartacea o all'e-mail) basata su solleciti telefonici ad opera di un *contact center* esterno o direttamente da parte dei rilevatori comunali, eventualmente mirata a specifici segmenti di popolazione.

4. Discussione e sviluppi futuri

L'esempio presentato nel paragrafo precedente è una semplice ma efficace descrizione di come le scelte da effettuare in fase di raccolta dati possano essere supportate da risultati quantitativi derivanti da modelli statistici che interpretano il processo di risposta in funzione di variabili ausiliarie note per tutta la popolazione di riferimento. Lo stesso tipo di analisi potrebbe ovviamente essere sfruttata a monte, ossia nella fase di progettazione di una nuova edizione di indagine, per migliorare quegli aspetti del processo che hanno influito negativamente sulla partecipazione all'indagine o per sperimentare l'uso di disegni adattivi, che sono finalizzati a ottenere una risposta più bilanciata individuando i sotto-processi della raccolta dati e i sottogruppi di unità con un impatto maggiore sulla rappresentatività della risposta (Bethlehem *et al*., 2011).

È possibile usare vari modelli statistici che, operando sotto condizioni diverse (Cobben, 2009), offrono differenti spunti di osservazione del processo di raccolta dati e, di conseguenza, suggeriscono l'adozione di azioni diverse, ma tutte mirate a effettuare quegli interventi correttivi che, nel migliorare il tasso di risposta in termini di entità e composizione, cercano anche di ottimizzare i costi e l'impegno di risorse umane dedicate alla conduzione e al monitoraggio della rilevazione. Rientrano in queste azioni anche i *responsive design* che hanno la stessa finalità dei disegni adattivi, ma, a differenza di questi, individuano i sotto-processi di raccolta dati e i

sottogruppi di unità in base alle informazioni raccolte via via durante la rilevazione stessa. Modelli logistici semplici o innestati, modelli a classi latenti, modelli *multilevel* o modelli di sopravvivenza sono alcuni esempi di applicazioni possibili finalizzate a identificare effetti di autocorrelazione nella risposta indotti dagli enti territoriali e/o dagli organi intermedi di rilevazione (comuni e rilevatori) o la presenza di una variabilità residua spiegabile da informazioni di cui non si sta tenendo conto.

Requisito fondamentale per la costruzione di qualsivoglia modello che vada nell'ottica descritta è, come già detto nel paragrafo 1, la disponibilità di paradati e variabili ausiliarie, ossia di quegli 'ingredienti' imprescindibili per l'analisi del processo di indagine. È quindi importante che l'organizzazione della raccolta dati sia tale da garantire la memorizzazione tempestiva e sistematica di queste informazioni e che queste vengano definite durante la progettazione dell'indagine stessa. Si pensi, ad esempio a tutte le informazioni relative alla conduzione di una rilevazione CAPI, come numero, giorno ed esito delle visite, oppure, in un'indagine CAWI, ai dati riguardanti la tempistica dei solleciti postali o via e-mail. Accanto a queste informazioni, vanno considerate anche tutte quelle presenti nei registri statistici disponibili in Istituto, quali ad esempio le informazioni demografiche, strutturali, sociali e territoriali, che rappresentano variabili note a monte del processo di raccolta dati e disponibili per tutta la popolazione di riferimento dell'indagine.

Se l'applicabilità dei modelli di analisi dipende dalla disponibilità di paradati e variabili ausiliarie, la loro usabilità dipende dal modo in cui vengono presentati i risultati. È, quindi, di cruciale importanza costruire strumenti di monitoraggio che permettano ai supervisori di indagine di interpretare in modo semplice e diretto i risultati dell'analisi, così da individuare facilmente quali modifiche apportare alla raccolta dati e quale sia l'impatto delle stesse sull'andamento della rilevazione.

Il grafico riportato in Figura 1 è una dimostrazione della semplicità auspicata, come potrebbero esserlo altri strumenti e indicatori di monitoraggio, come, ad esempio, le carte di controllo e/o gli R-*indicators* (RISQ, 2009), che permettono di supervisionare più aspetti del processo di raccolta dati e, di conseguenza, consentono di intervenire in modo rapido sul campo con azioni mirate e specifiche per il problema incontrato. Ad esempio, le carte di controllo potrebbero essere usate per monitorare il comportamento del rilevatore CATI durante la fase di contatto per capire se viene registrato un numero di rifiuti troppo alto rispetto alla media (Murgia e Simeoni, 2005), posizionandosi costantemente e non in modo sporadico al di sopra del limite superiore della carta di controllo. In tal caso, si potrebbe pensare di intervenire con un *debriefing* sul come motivare i rispondenti alla partecipazione. Gli R-*indicators*, a loro volta, potrebbero essere usati per individuare i segmenti di popolazione sottorappresentati in modo da agire su di essi con solleciti mirati o assegnandoli a rilevatori più esperti.

L'uso di modelli di analisi della non-risposta e di strumenti di monitoraggio deve essere inquadrato all'interno di una strategia per la qualità dell'intero processo di indagine e non limitata alla sola fase di conduzione della raccolta dati. Tale strategia dovrà essere finalizzata a: i) definire obiettivi di qualità misurabili, ii) scegliere i modelli di analisi più adatti a individuare, per poi correggere, le cause distorsive della non-risposta; iii) rendere disponibili i paradati e le variabili ausiliarie di interesse; iv) costruire strumenti di monitoraggio semplici e applicabili a tutte le indagini, a prescindere dalla popolazione e dalla tecnica di raccolta dati usata.

La strategia dovrà essere articolata nei seguenti passi:

1) Scelta del modello e degli strumenti. Per individuare i modelli e gli strumenti di monitoraggio più adatti agli scopi prefissati, è necessario partire dai dati di indagini già concluse e per le quali si hanno informazioni sulle azioni messe in campo durante la fase di raccolta dati. Da questa prima fase della strategia si potrà non solo valutare l'efficacia e l'applicabilità dei metodi, ma anche stabilire se le azioni di *field* messe effettivamente in atto abbiano contribuito a migliorare la rappresentatività della risposta. Per questo ultimo aspetto ci si potrebbe basare sulla definizione di scenari derivanti da modelli che usano dati simulati.
2) Sperimentazione dei modelli e degli strumenti. Una volta individuati i metodi e gli strumenti, questi potrebbero essere usati in via sperimentale su indagini che devono andare in *field* e per le quali si dispone di variabili ausiliarie. La sperimentazione potrebbe essere fatta in corso di indagine oppure durante un'indagine pilota. Nel primo caso, dal campione di indagine potrebbe essere estratto un sotto-campione casuale di unità che a sua volta potrebbe essere diviso in due gruppi: uno di test e uno di controllo (RISQ, 2009). Al sottogruppo di test si applicano i nuovi strumenti di monitoraggio e le azioni correttive da questi suggerite, mentre sul sottogruppo di controllo si opera come di consueto. Al termine della rilevazione si potrà così osservare se il campione di test ha ottenuto risultati migliori in termini di bilanciamento della mancata risposta.
3) Definizione, raccolta e memorizzazione di paradati e variabili ausiliarie. A monte delle sperimentazioni, e in modo sistematico per tutte le indagini, dovrà essere fatto uno studio per capire quali informazioni ausiliarie e paradati sono già disponibili e quali sono quelle da aggiungere tramite collegamento a registri e/o progettando la raccolta di paradati attraverso i questionari elettronici, per le indagini che fanno uso di tecniche assiste da computer, e mediante l'osservazione del rilevatore per le indagini dirette. In caso di rilevazione affidata a ditte esterne, sarà importante prevedere la raccolta, memorizzazione e invio di queste informazioni in modo sistematico e tempestivo durante la fase di raccolta dati. A latere si potrebbe perseguire un obiettivo di medio-lungo periodo che è quello di semplificare la quantità di reportistica di monitoraggio sviluppata dalle ditte

esterne che genera sovente un carico di lavoro eccessivo per entrambi le parti. Per le indagini il cui questionario elettronico è sviluppato in Istituto, dovranno essere date indicazioni sul tipo di paradati da memorizzare.

4) Centralizzazione dei sistemi di monitoraggio. Dovranno essere potenziate e sfruttate al massimo le funzionalità offerte dall'esistenza di un sistema centralizzato per la gestione della fase di raccolta dati in modo da standardizzare la rilevazione delle informazioni per tipologia di indagine e facilitare così l'implementazione e l'utilizzo dei modelli di analisi della non-risposta.

A monte della strategia sopra decritta è fondamentale tenere conto del disegno del questionario di indagine che deve essere tale da contenere l'errore di misura dovuto alla tecnica (*mode measurement effect*). Ciò è vero soprattutto nelle indagini *mixed-mode*, per le quali questo errore può verificarsi quando uno stesso rispondente risponde diversamente ad uno stesso quesito al variare della tecnica di rilevazione utilizzata, e inoltre perché non tutte le tecniche sono 'adatte' a somministrare qualsiasi tipologia di domanda (Istat, 2018). In questi casi è importante che la fase di raccolta dati sia preceduta da una fase di test del questionario mirata a scegliere il disegno che limiti l'impatto distorsivo del *mode measurement effect*, nel caso valutando anche se la misurazione di una o più variabili target possa essere influenzata dalla tecnica, indipendentemente dalla diversa propensione ad essa di particolari sottoinsiemi di rispondenti.

Riferimenti bibliografici

COBBEN F. 2009. Nonresponse in household surveys. Methods for analysis and adjustment. PhD thesis. University of Amsterdam, Statistics Netherlands, Amsterdam.

BETHLEHEM J.G., COBBEN F., SCHOUTEN B. 2011. *Handbook of nonresponse in household surveys*. Wiley Handbooks in Survey Methodology.

ISTAT (2018). L'effetto tecnica nelle indagini mixed-mode – Aspetti teorici e sperimentazioni su indagini sociali che utilizzano il web. Collana Letture statistiche: Metodi. Disponibile al link: https://www4.istat.it/it/archivio/211135.

KREUTER F. (Ed.). 2013. *Improving surveys with paradata: Analytic uses of process information* (Vol. 581). Hoboken: John Wiley & Sons.

LITTLE R.J.A., RUBIN D.B. 2002. *Statistical analysis with missing data.* Wiley series in Probability and Statistics. New York: John Wiley & Sons, Inc., NY, USA.

LYBERG L.E., STUKEL D.M. 2017. The roots and evolution of the total survey error concept. In BIEMER P.P et al. (Eds.), *Total Survey Error in Practice*, John Wiley & Sons, pp. 1-22.

MURGIA M., SIMEONI G. 2005. Improving the quality of assisted coding of occupation in CATI surveys through control charts. CLADAG - Classification and Data Analysis Group 2005, 6-8 giugno, Parma.
RISQ – Representative Indicators for Survey Quality (2009). Disponibile al link: https://www.cmi.manchester.ac.uk/research/projects/representative-indicators-for-survey-quality/publications/.

SUMMARY

Quality in data capturing processes: the case of the survey on the aspects of daily life

The quality of the statistical information depends on several factors such as definitions and concepts adopted, sampling design, data capturing techniques, tools and methods for the analysis and representations of data. The Total Survey Error (TSE) approach represents a theoretical framework for the evaluation of the quality of survey data that takes into account the different types of error that might arise in each step of the survey process, from the design phase to the analysis of data, in order to compare costs and benefits. In this framework, the survey error is defined as the deviation of a survey response from its underlying true value. For every phase of the survey it is necessary to evaluate the efficiency of the actions taken. This work focuses on the assessment of the data collection phase through indicators based on data that are collected during the process itself, the so-called paradata.

The analysis of paradata and their association with other information -auxiliary data as well as survey data- is useful in constructing performance indicators of the data collection phases to guide the necessary adjustment process.

This paper reports an example of the approach followed, based on the data of the Istat survey 'Aspects of daily life, 2019'. Specifically, it describes the analysis of the propensity to respond and the data collection techniques used (CAWI vs CAPI-PAP) to detect those respondents profiles that need an adaptation of either contact or reminder strategies to improve their propensity to respond. In this example the analysis is conducted using a logistic model. The possible methods and tools for improving the quality of the survey are drew as part of a strategy involving the different actors of the data collection process.

Claudio CECCARELLI, Istat, clceccar@istat.it
Marco FORTINI, Istat, fortini@istat.it
Manuela MURGIA, Istat, murgia@istat.it
Alessandra NUCCITELLI, Istat, nuccitel@istat.it
Rita RANALDI, Istat, ranaldi@istat.it
Francesca ROSSETTI, Istat, frrosset@istat.it