

istat working papers

N.16
2015

Integrazione da fonti amministrative per il calcolo delle stime campionarie: il caso dell'Indagine sui consumi energetici delle famiglie

Simona Rosati, Valentina Talucci

istat working papers

N.16
2015

Integrazione da fonti amministrative per il calcolo delle stime campionarie: il caso dell'Indagine sui consumi energetici delle famiglie

Simona Rosati, Valentina Talucci

Comitato scientifico

Giorgio Alleva
Tommaso Di Fonzo
Fabrizio Onida

Emanuele Baldacci
Andrea Mancini
Linda Laura Sabbadini

Francesco Billari
Roberto Monducci
Antonio Schizzerotto

Comitato di redazione

Alessandro Brunetti
Romina Fraboni
Maria Pia Sorvillo

Patrizia Cacioli
Stefania Rossetti

Marco Fortini
Daniela Rossi

Segreteria tecnica

Daniela De Luca Laura Peci Marinella Pepe Gilda Sonetti

Istat Working Papers

Integrazione da fonti amministrative per il calcolo delle stime campionarie: il caso dell'Indagine sui consumi energetici delle famiglie

N. 16/2015

ISBN 978-88-458-1849-3

© 2015

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Salvo diversa indicazione la riproduzione è libera,
a condizione che venga citata la fonte.

Immagini, loghi (compreso il logo dell'Istat),
marchi registrati e altri contenuti di proprietà di terzi
appartengono ai rispettivi proprietari e
non possono essere riprodotti senza il loro consenso.

Integrazione da fonti amministrative per il calcolo delle stime campionarie: il caso dell'Indagine sui consumi energetici delle famiglie¹

Simona Rosati², Valentina Talucci³

Sommario

Le indagini CATI, sebbene presentino molti vantaggi, tra cui la riduzione dei costi, un'elevata tempestività e una maggiore semplificazione nella gestione della rete di rilevazione, risentono degli effetti dovuti all'errore di sottocopertura delle liste telefoniche. Ciò, come è noto, si ripercuote sulla bontà degli stimatori, che possono risultare anche molto distorti. Questo lavoro ha lo scopo di evidenziare gli aspetti metodologici che caratterizzano la strategia di stima, nell'ambito dell'Indagine sui consumi energetici delle famiglie, al fine di ridurre la distorsione causata dalle imperfezioni della lista di campionamento. L'elemento innovativo di questa fase è costituito dall'utilizzo di informazioni ausiliarie altamente correlate con la variabile di interesse, desunte da archivi di fonte amministrativa. Tali informazioni sono state integrate nel campione e poi inglobate nel procedimento di stima.

Parole chiave: CATI, sottocopertura, stimatore calibrato, dati amministrativi.

Abstract

Although CATI surveys have many advantages, such as cost reduction, high timeliness and simpler management of the interviewer network, they are affected by under-coverage errors, which affect the residential phone directory. This, in turn, affects the quality of estimators, which may be highly inaccurate due to large bias. The aim of this work is to stress the methodological features of the estimation strategy adopted for the Household energy consumption survey, in order to reduce the bias caused by the imperfection of the sampling frame. The innovative element of this phase is the use of auxiliary information from administrative data sources which are highly correlated with the target variables. This information was integrated in the sample and then included in the estimation process.

Keywords: CATI, under-coverage, calibration estimator, administrative data.

¹ Una versione preliminare di questo lavoro è stata presentata alle Giornate della ricerca in Istat del 10-11 novembre 2014. Le opinioni espresse in questo lavoro impegnano esclusivamente gli autori e non implicano alcuna responsabilità da parte dell'Istat.

² DISA/A – Innovazione di processo e di prodotto – email: sirosati@istat.it

³ AMB – Servizio Stato dell'ambiente – email: talucci@istat.it

1. Introduzione

Nel 2011 l'Agencia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile (ENEA) e l'Istituto nazionale di statistica (Istat) hanno stipulato un 'Accordo di collaborazione per la progettazione e la realizzazione di un'indagine sui consumi energetici del settore residenziale'. Tale indagine, realizzata tra marzo e giugno del 2013 attraverso una tecnica CATI (*Computer Assisted Telephone Interviewing*), ha tra i suoi obiettivi quello di stimare i consumi energetici e le spese delle famiglie per destinazione finale (cucina, riscaldamento e raffrescamento degli ambienti, illuminazione, utilizzo degli elettrodomestici) e fonte energetica (gas, energia elettrica, gasolio, biomasse, ecc.).

L'obiettivo di questo lavoro è dare risalto alle innovazioni di processo e di prodotto che hanno portato a un significativo miglioramento delle stime. Partendo dal presupposto che le stime sono tanto più precise quanto più le variabili ausiliarie sono correlate con la variabile di interesse, è stata individuata la variabile 'reddito familiare complessivo' quale ulteriore variabile ausiliaria altamente correlata con il fenomeno oggetto di stima. Tale variabile, desunta dall'archivio amministrativo dei redditi, è stata integrata nel campione e poi inglobata nel procedimento di stima. Ciò ha rappresentato un valido metodo di utilizzo di dati amministrativi integrati per contrastare gli effetti distorsivi dovuti alle imperfezioni della lista di campionamento.

2. Il campione

Il disegno di campionamento è a uno stadio, con stratificazione dei circa 8.000 comuni italiani per regione e tipologia socio-demografica dei comuni (definita in base a ampiezza demografica e zona altimetrica). La procedura di selezione delle famiglie è casuale semplice, a partire dall'archivio informatizzato ufficiale delle famiglie abbonate alla rete di telefonia fissa, costantemente aggiornato. Successivamente all'interno di ciascuna famiglia sono stati selezionati i rispondenti tra gli individui eleggibili di età superiore ai 18 anni, in quanto indicati dalle famiglie stesse come i più idonei a fornire informazioni sui fenomeni oggetto di intervista.

Al fine di garantire il numero di interviste programmate, pari a 20.000, è stato estratto un campione di famiglie sostituite pari a cinque volte quello base, per un totale di 20.000 sestine ovvero 120.000 unità complessive. Meccanismi di sostituzione per gli individui erano previsti anche all'interno della famiglia stessa. Ciò ha consentito di ottenere un campione osservato di numerosità pari a quello estratto.

3. La qualità delle stime nelle indagini CATI

Se da un lato l'utilizzo della tecnica CATI presenta indubbi vantaggi, tra i quali una gestione più semplificata della rete di rilevazione, il contenimento dei costi e la tempestività della raccolta dei dati, ciò non è, di per sé, una garanzia di qualità. Essa infatti necessita comunque di essere monitorata e i suoi possibili effetti devono essere valutati (Istat 2005).

A queste considerazioni vanno ad aggiungersi quelle riguardanti gli errori di copertura delle liste telefoniche, che nel corso del tempo hanno assunto sempre più rilevanza a causa delle loro importanti implicazioni per le indagini condotte con tecnica CATI.

È noto, infatti, come nel corso degli anni la lista ufficiale delle famiglie abbonate alla rete di telefonia fissa, che viene utilizzata per selezionare i campioni delle indagini CATI, abbia risentito di una progressiva diminuzione del tasso di copertura. In Italia si assiste ad un calo progressivo della quota di famiglie in possesso del telefono fisso, passata dal 72,8 per cento del 2009 al 64,0 per cento del 2013, a cui si contrappone la crescita del possesso del telefono cellulare, che passa dall'87,8 per cento del 2009 al 91,5 per cento del 2013 (Istat 2014).

Si pone dunque il problema del grado di rappresentatività dei campioni estratti dalle liste telefoniche a causa della sottocopertura e in definitiva della probabile distorsione delle stime, qualora le

unità appartenenti alla lista rivelino un comportamento significativamente diverso da quelle non incluse nella lista rispetto alle variabili di interesse dell'indagine.

4. Lo stimatore

Lo stimatore utilizzato per il calcolo dei coefficienti di riporto all'universo è lo *stimatore di ponderazione vincolata*, metodo largamente utilizzato nell'ambito della statistica ufficiale. Lo stimatore di ponderazione vincolata, noto in letteratura come *calibration estimator* (Deville, Särndal 1992; Särndal 2007), oltre a migliorare l'accuratezza delle stime, ha il vantaggio di garantire la coerenza delle stime prodotte con le informazioni ausiliarie note (totali di popolazione noti da fonti esterne all'indagine). Questa classe di stimatori comporta la costruzione di pesi di riporto all'universo che sono molto vicini a quelli base (o da disegno), ma che contemporaneamente soddisfano vincoli di coerenza sulle variabili ausiliarie, imposti attraverso le cosiddette equazioni o vincoli di calibrazione.

Inizialmente le stime sono state calcolate prendendo in considerazione i seguenti totali noti:

- popolazione residente per sesso e classe di età nelle cinque ripartizioni territoriali (Nord Est, Nord Ovest, Centro, Sud e Isole);
- popolazione residente per regione (incluse Trento e Bolzano);
- numero di famiglie residenti per regione;
- popolazione di 15 anni e oltre per condizione professionale e posizione nella professione (lavoratori alle dipendenze, autonomi, disoccupati, altri).

I primi tre totali sono stati desunti da fonti demografiche (anagrafiche), mentre i totali riferiti alla condizione professionale e alla posizione nella professione derivano dall'Indagine sulle forze di lavoro (anno 2013). Indicheremo con T_1 lo stimatore corrispondente.

Non essendo soddisfacenti i risultati ottenuti con T_1 , presumibilmente a causa degli effetti distorsivi dovuti alle imperfezioni della lista di campionamento, è stato necessario individuare ulteriori variabili ausiliarie correlate con la variabile oggetto di interesse, con lo scopo di ridurre la distorsione e allo stesso tempo migliorare la precisione delle stime. Recenti studi dimostrano infatti come un approccio di questo tipo possa portare a un significativo miglioramento della riduzione della distorsione dovuta alla sottocopertura della lista (De Vitiis, Righi 2011).

È stata così individuata la variabile 'reddito familiare complessivo' desumibile dall'archivio della Banca dati reddituale che è stata opportunamente integrata nel campione e poi inglobata nel procedimento di stima. Chiameremo T_2 lo stimatore corrispondente. L'ipotesi che si vuole avvalorare è che, essendo il reddito altamente correlato con la variabile oggetto di studio, lo stimatore T_2 produrrà stime più accurate rispetto a T_1 .

Lo stimatore T_2 può essere costruito essenzialmente in due diversi modi a seconda dell'utilizzo della variabile ausiliaria 'reddito'. Un primo metodo, il più immediato, consiste nel definire ulteriori post-strati che tengano conto anche del 'reddito familiare complessivo'; in alternativa, la stessa variabile può essere impiegata per definire opportune classi di aggiustamento analogamente a quanto avviene per le mancate risposte⁴. Tale metodo può costituire un metodo per la correzione dell'errore di sottocopertura della lista di campionamento, nella misura in cui sono note a livello di popolazione obiettivo le dimensioni delle classi di aggiustamento e vale l'ipotesi che la media delle unità listate in ciascuna classe è uguale a quella delle unità non listate (Nicolini *et al.* 2013). Ed è pertanto in base a tali considerazioni che è stato adottato il metodo delle classi di aggiustamento.

Le classi o celle di aggiustamento sono state costruite riportando per ciascuna delle cinque ripartizioni territoriali la corrispondente distribuzione dei quintili di reddito; quindi, sulla base di quanto

⁴ Il metodo delle classi di aggiustamento è tanto più efficace quanto più è valida l'ipotesi su cui si basa, e cioè la capacità delle variabili utilizzate di definire classi caratterizzate da un diverso tasso di risposta, ma per le quali la media del carattere tra i rispondenti è uguale a quella dei non rispondenti. La motivazione sta nel fatto che la distorsione da mancata risposta potrebbe essere anche annullata con un'opportuna riponderazione dei dati, cioè con una modifica dei pesi di riporto all'universo attribuiti ai valori osservati sui soli rispondenti (Nicolini *et al.* 2013).

è noto per l'intera popolazione, per ciascuna cella è stato calcolato un fattore correttivo tale che, moltiplicato per il peso base, riproducesse nel campione la distribuzione del reddito della popolazione. Si usa in questo caso parlare anche di *riponderazione* dei dati, intendendo con tale termine la modifica dei pesi base al fine di incrementare l'efficienza dello stimatore e in senso lato la rappresentatività delle unità campionarie mediante il loro peso di riporto.

5. Primi risultati

Riportando il grafico a dispersione delle coppie dei coefficienti di riporto all'universo, ottenuti con i due stimatori T_1 e T_2 (Figura 1), emerge chiaramente una maggiore variabilità dello stimatore T_2 (post aggiustamento) rispetto a T_1 (pre aggiustamento), come del resto era prevedibile, avendo introdotto un'ulteriore variabile nel processo di stima. Tale risultato è comprovato dai coefficienti di variazione calcolati sulle stime prodotte (120 contro 135), sebbene gli stessi indicatori calcolati per i correttori dei pesi base, pre e post aggiustamento, siano più contenuti per questi ultimi, a dimostrazione del fatto che la correzione per celle di aggiustamento produce stime iniziali meno distanti da quelle finali di calibrazione (Tavola 1).

Figura 1 – Confronto tra pesi finali pre e post aggiustamento

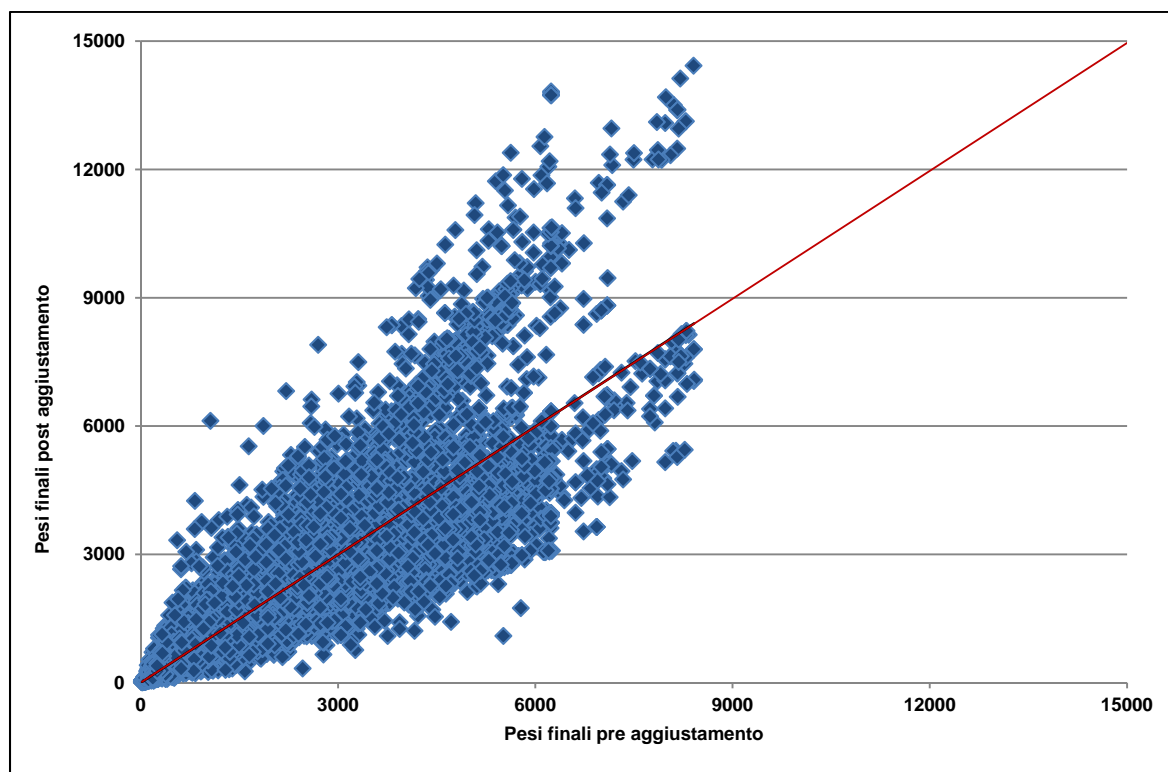
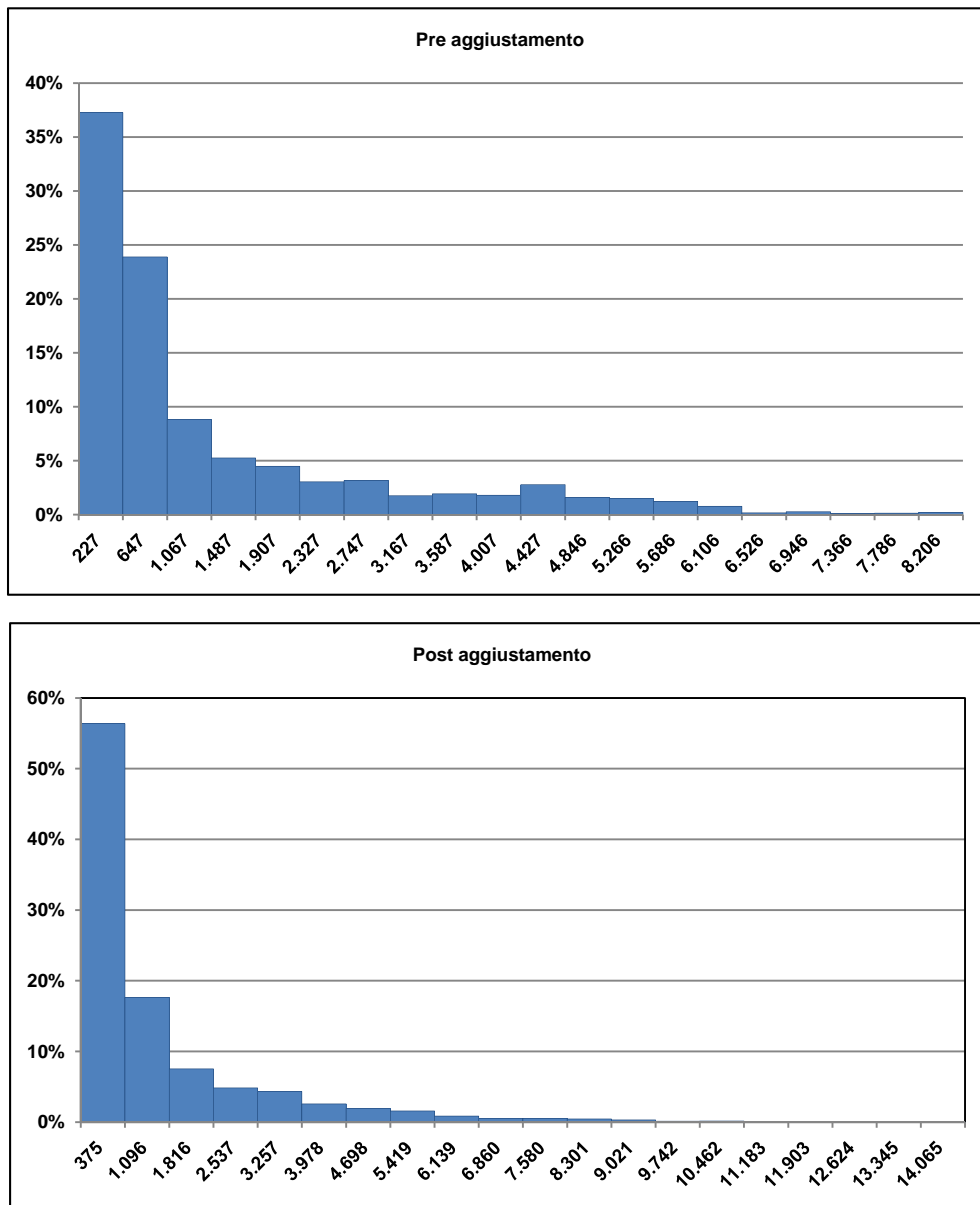


Tavola 1 – Coefficiente di variazione dei pesi finali e dei correttori dei pesi diretti – livello ripartizionale

Coefficiente di variazione	Nord-ovest	Nord-est	Centro	Sud	Isole
Pesi finali					
pre aggiustamento	113	114	114	128	114
post aggiustamento	124	125	133	146	135
Correttori dei pesi diretti					
pre aggiustamento	87	81	92	97	104
post aggiustamento	83	76	88	91	102

Le distribuzioni dei pesi finali pre e post aggiustamento mostrano anch'esse un andamento differente: più regolare e omogenea quella ottenuta con il metodo delle celle di aggiustamento (Figura 2). Per motivi di esposizione riportiamo soltanto le distribuzioni complessive, ma si può far vedere come la stessa analisi condotta per ciascuna ripartizione territoriale renda ancora più evidente questo risultato. Ciò significa che lo stimatore T_2 produrrà stime maggiormente stabili rispetto a T_1 nei diversi domini di studio.

Figura 2 – Distribuzione dei pesi finali pre e post aggiustamento – livello nazionale



6. La fase di integrazione con gli archivi amministrativi

Sempre più frequentemente, nella produzione delle statistiche ufficiali, gli Istituti nazionali di statistica affiancano alla raccolta diretta dei dati basata su indagini censuarie o campionarie l'impiego di informazioni provenienti da archivi amministrativi, allo scopo di migliorare il processo di produzione statistica.

In ogni processo di produzione statistica la qualità dei risultati prodotti dipende largamente dalla

qualità dei dati di *input*. Analogamente a quanto accade per le indagini dirette, l'uso dei dati di fonte amministrativa⁵ implica una fase di valutazione dei requisiti di qualità necessari affinché la fonte stessa possa essere ritenuta effettivamente utilizzabile a fini statistici. La valutazione della qualità riveste, dunque, un ruolo rilevante sia nel caso di fonti già inserite nel processo produttivo, che necessitano di essere costantemente monitorate al fine di rilevare errori imprevedibili o variazioni connesse, ad esempio, a eventuali cambiamenti normativi, sia per le nuove fonti al fine di verificare la loro utilizzabilità nel processo di produzione. Quando un archivio amministrativo viene acquisito da un Istituto nazionale di statistica risulta pertanto necessario che esso sia affiancato da una serie di indicatori, utili a fornire una prima indicazione della sua qualità⁶ (Daas 2011).

Solo una volta chiusa questa prima fase o "ciclo di vita" dei microdati è possibile passare alla seconda fase di integrazione dei dati da fonti diverse.

L'obiettivo di questo lavoro è quello di ricavare la variabile ausiliaria (X) "reddito familiare complessivo" attraverso un processo di integrazione di più fonti amministrative.

Le fonti amministrative individuate sono due: la prima è la Banca Dati Reddittuale di fonte Ministero dell'Economia e Finanza, che conta circa 41 milioni di *record*, in cui è contenuta la variabile ausiliaria di interesse 'reddito' e si compone di tre gruppi di *file* relativi ai contribuenti che hanno compilato: il modello UNICO Persone fisiche⁷ o il modello 730⁸ o le dichiarazioni presenti solo nei modelli 770⁹.

Il reddito complessivo è quello che viene utilizzato per determinare l'imposta sui redditi delle persone fisiche (Irpef), dovuta per l'anno d'imposta; è dato dalla somma dei singoli redditi: dominicali, agrari, da fabbricati, da lavoro dipendente o autonomo, d'impresa in contabilità ordinaria, d'impresa in contabilità semplificata, d'impresе consorziate, da partecipazione, da plusvalenze di natura finanziaria, altri redditi, da allevamento, da tassazione separata.

La seconda fonte utilizzata per l'integrazione è la Lista Anagrafica Comunale (LAC), che conta circa 24 milioni di famiglie e che consente di connettere ogni unità individuale alla rispettiva famiglia anagrafica.

Il processo d'integrazione fa riferimento a metodologie di *record linkage* che possono essere di tipo deterministico, quando si dispone di chiavi univoche di aggancio, o probabilistiche quando si ricorre a profili di similarità tra unità ottenuti da un insieme di variabili caratterizzanti i *record*. Lo scopo di un processo di integrazione è individuare la presenza della stessa unità elementare (persona fisica, persona giuridica, luogo, relazione) nelle diverse fonti prescelte; questo obiettivo è reso possibile attribuendo una o più chiavi di identificazione univoca e stabile nel tempo.

Nel caso dell'Indagine sui consumi energetici delle famiglie (*data set target* Y), l'integrazione tra le diverse fonti individuate è stata effettuata attraverso una metodologia di *record linkage* di tipo deterministico, resa possibile grazie alla presenza di chiavi di aggancio univoche, cioè di una o più variabili comuni per tutte le fonti.

Come si può apprezzare dal Prospetto 1, il processo di integrazione è stato preceduto da una fase "zero" di organizzazione e pretrattamento che ha consentito di riconnettere ad ogni singola unità campionaria la variabile chiave 'codice fiscale' proveniente dall'archivio dei numeri telefonici (Consodata).

⁵ Per dati di fonte amministrativa si intendono dati derivati da una o più fonti amministrative, prima di ogni processo di validazione statistica; una fonte amministrativa è un insieme di dati raccolti e mantenuti (da un soggetto terzo rispetto all'unità cui i dati si riferiscono) per l'attuazione di uno o più regolamenti amministrativi; infine una fornitura amministrativa si riferisce ad uno o più dataset amministrativi ricevuti dall'Istituto nazionale di statistica da parte del Fornitore di dati amministrativi in seguito ad una specifica richiesta.

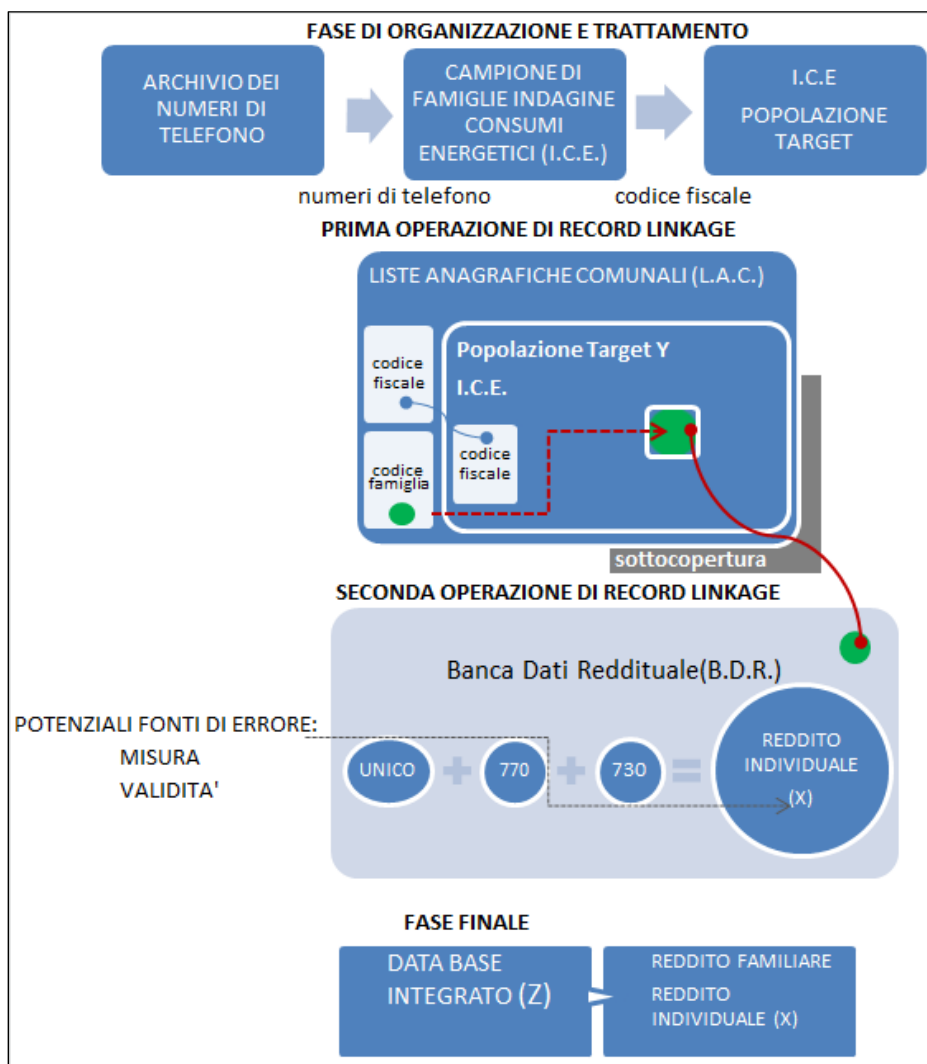
⁶ Questo implica evidentemente la definizione del concetto di qualità per questo tipo di dati. Facendo riferimento ai più recenti *framework* teorici che l'Europa fornisce in materia di qualità dei dati di input di fonte amministrativa, essa viene misurata declinando cinque dimensioni considerate essenziali: Controlli tecnici (*Technical checks*), Integrabilità (*Integrability*), Accuratezza (*Accuracy*), Completezza (*Completeness*), Dimensione temporale (*Time-related dimension*).

⁷ In particolare il modello UNICO è un modello unificato che permette di presentare più dichiarazioni fiscali. Devono compilare la dichiarazione in forma unificata i contribuenti tenuti alla presentazione sia della dichiarazione dei redditi sia della dichiarazione Iva (sono escluse alcune categorie di contribuenti chiamati a presentare la dichiarazione Iva in forma autonoma).

⁸ I lavoratori dipendenti e i pensionati (in possesso di determinati redditi) possono presentare la dichiarazione con il modello 730; in alcuni casi è possibile presentare il modello 730 anche in assenza di un datore di lavoro tenuto a effettuare il conguaglio.

⁹ Il modello 770, ordinario o semplificato, viene utilizzato dai sostituti d'imposta per comunicare i dati fiscali relativi alle ritenute operate.

Prospetto 1 – Le fasi dell’integrazione tra dati amministrativi e dati campionari



Si è quindi effettuato un primo *linkage* deterministico tra i *record* del campione e l’archivio delle liste anagrafiche comunali (LAC) attraverso la chiave identificativa univoca ‘codice fiscale’; il tasso di abbinamento è stato del 91%, questo significa che per poco più di 9 famiglie su 10 è possibile disporre delle chiavi di aggancio con la Base Dati dei Redditi (BDR). In questa fase ogni errore nei dati di *input* o incompatibilità fra dati può dare origine a un errore di copertura della popolazione *target*, in particolare di sottocopertura, cioè individui presenti nel campione e non nella LAC. La causa di questa sottocopertura può essere addebitata ad errori formali nella chiave di aggancio ‘codice fiscale’.

Sono possibili anche errori di allineamento. Lo scopo dell’allineamento è di identificare tutte le relazioni fra le unità abbinate (individui) per la creazione dell’unità statistica (famiglia anagrafica). Nel caso oggetto di studio si tratta della relazione che lega ogni singolo soggetto, identificato dal codice fiscale, con la rispettiva famiglia anagrafica identificata dal codice famiglia. Una relazione di questo tipo viene detta “di molti a uno” fra le unità di base e le unità composte. Fonti diverse, in questo caso LAC e BDR, possono contenere informazioni in conflitto relativamente alle relazioni fra unità di base e unità composte, cioè un individuo può far parte di un “grappolo” famiglia in un archivio e non nell’altro, a causa di variazioni anagrafiche, mancate cancellazioni, ecc..

La seconda fase di *linkage*, anche in questo caso di tipo deterministico, è stata effettuata tra gli individui del campione, dove è stato possibile assegnare la chiave identificativa della famiglia co-

me concatenamento del codice comune, codice provincia e codice famiglia anagrafica, e la Banca Dati Reddittuale. L'*output* così ottenuto ha consentito di attribuire per ciascuna famiglia del campione, che si è agganciata con la BDR, il reddito complessivo familiare, che si ricorda essere stato calcolato come somma dei redditi individuali. Il tasso di abbinamento di questa seconda fase è stato superiore al 90%.

Si è ottenuto così il *data set* integrato (Z), a partire dalla popolazione *target* (Y), sul quale si è importata la variabile ausiliaria 'reddito familiare' (X). Se le funzioni di distribuzione del reddito sulla popolazione *target* $f(Y=y, X=x)$ e sul *data set* integrato $f(Z=y, X=x)$ hanno la stessa distribuzione vuol dire che i dati integrati (Z, X) forniscono inferenza valida se usati al posto di (Y, X). La validità statistica è così definita rispetto alle funzioni di distribuzione (Zhang 2012).

7. Conclusioni e ulteriori sviluppi

Abbiamo visto come l'utilizzo di una variabile altamente correlata con la variabile di interesse contribuisca, in fase di aggiustamento dei pesi base, a ridurre la distanza tra questi e i pesi finali, pur aumentando la variabilità. Tuttavia è bene sottolineare che una maggiore variabilità dello stimatore non necessariamente implica una minore accuratezza dello stesso, in quanto tale aumento potrebbe essere più che compensato da una minore distorsione dello stimatore e di conseguenza anche l'errore totale potrebbe diminuire.

Per valutare l'effettiva riduzione della distorsione dello stimatore si può ricorrere a metodi indiretti, essendo il valore del parametro non noto. Come suggerito da altri autori si può fare riferimento all'indagine 'Aspetti della vita quotidiana' – AVQ (De Vitiis, Righi 2011). Il campione dell'indagine AVQ, infatti, è selezionato dalle liste anagrafiche comunali e tra le variabili rilevate include sia quella specifica sulla presenza o meno di una linea telefonica fissa nell'abitazione sia altre riconducibili a diverse indagini ufficiali, tra cui quella in esame. L'idea è quella di confrontare, a parità di stimatore, le stime ottenute rispetto sia al campione osservato sia al sottoinsieme delle famiglie con telefono. Pertanto una divergenza significativa tra le stime sarebbe imputabile a una distorsione dovuta alla distribuzione della sottopopolazione delle famiglie con telefono. Applicando lo stesso procedimento per entrambi gli stimatori, T_1 e T_2 , dove il secondo tiene conto dell'informazione aggiuntiva sul reddito complessivo familiare, ci si attende che questo contribuisca a ridurre la distorsione.

A completamento dell'analisi fin qui svolta si vuole procedere con la valutazione della strategia di stima tramite gli errori di campionamento, l'effetto del disegno (*deft*) e l'effetto stimatore; infine, si vuole fornire una misura di dispersione, che scompone il cambiamento totale tra i pesi da disegno e i pesi finali nel contributo introdotto dalla correzione iniziale, dalla calibrazione finale e dall'interazione tra i due passi (Dufour *et al.* 2001).

Riferimenti bibliografici

- Daas, P., S. Ossen, M. Tennekes, L.C. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren e B. Wallgren. *Report on methods preferred for the quality indicators of administrative data sources*. Second deliverable of WP4 of the BLUE Enterprise and Trade Statistics project. 2011. September 28. <http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf>
- De Vitiis, C. e P. Righi. Evaluations on list undercoverage bias and possible solutions: the case of ISTAT CATI survey “Trips, holidays and daily life”. 2011. *Rivista di Statistica Ufficiale*. 2-3: 5-19.
- Deville, J.C. and C.E. Sarndal. Calibration Estimators in Survey Sampling. 1992. *Journal of american statistical association*. 87 (418): 376-382.
- Dufour, J., F. Gagnon, Y. Morin, M. Renaud and C.E. Särndal. A Better Understanding of Weight Transformation Through a Measure of Change. 2001. *Survey methodology*. 27 (1): 97-108.
- Istat. *Annuario statistico italiano*. 2014. <http://www.istat.it/it/files/2014/11/Asi-2014.pdf>
- Nicolini, G., D. Marasini, G.E. Montanari, M. Pratesi, M.G. Ranalli e E. Rocco. *Metodi di stima in presenza di errori non campionari*. 2013. Springer. UNITEXT-Collana di Statistica e Probabilità Applicata.
- Sarndal, C.E. The calibration approach in survey theory and practice. 2007. *Survey methodology*. 33 (2): 99-119.
- Zhang, L.C. Topics of statistical theory for register based statistics and data integration. 2012. *Statistica Neerlandica*. 66 (1): 41-63.