L’indagine Istat sui consumi energetici delle famiglie: aspetti metodologici

Carlo Lucarelli
Istat - Direzione centrale delle statistiche socio-demografiche e ambientali
carlo.lucarelli@istat.it

Claudio Ceccarelli
Istat - Dipartimento per le statistiche sociali ed ambientali
claudio.ceccarelli@istat.it

Roma, 15 Dicembre 2014 – «I consumi energetici delle famiglie»

Indice

1. La rilevazione
2. Il controllo e il trattamento dei dati
3. La validazione con le altre fonti disponibili
4. La strategia di campionamento - disegno
5. La strategia di campionamento – stima
6. La strategia di campionamento – lo stimatore finale

La rilevazione

TEST DEL QUESTIONARIO: *svolto in diverse fasi, attraverso 3 tornate di pre-test svolti da ricercatori Istat su campioni ragionati di intervistati. Ha portato all'adozione di un allegato da compilare prima delle interviste per i quesiti più critici (consumi e spese, presenza di isolamento termico, ecc.).*

INDAGINE PILOTA*: Marzo-Giugno 2012 (1000 interviste con tecnica CATI con lettera di preavviso con allegato solo per metà del campione).*

RILEVAZIONE*: 27 Marzo 2013 - 29 Luglio 2013 (interviste con tecnica CATI con lettera di preavviso con allegato). 20.000 famiglie (campione base) + 100.000 famiglie (campione di riserva*).

PROCEDURA DI SELEZIONE DELLE FAMIGLIE: *casuale semplice da archivio informatizzato ufficiale delle famiglie abbonate alla rete di telefonia fissa aggiornato a Gennaio 2013.*

La rilevazione

RISPONDENTI: *selezionati tra gli individui eleggibili all'interno della famiglia (età >18 anni), indicati dalle famiglie stesse come i più idonei a fornire informazioni.*

NUMERO INTERVISTATORI FORMATI*: 150 circa (133 operativi)*

6 SESSIONI FORMATIVE: *dal 18 Marzo al 12 Aprile) della durata di 3 giornate (2 teoriche; 1 briefing tecnico con interviste di prova) + debriefing prima di avvio rilevazione.*

MONITORAGGIO*:*

- ✓ *Tramite indicatori di qualità.*
- ✓ *Sul campo (presenza di personale Istat in sala interviste per tutta la durata della rilevazione).*

DURATA MEDIA DELL'INTERVISTA*: 24 minuti circa.*

Il controllo e il trattamento dei dati

Controllo e correzione durante l'intervista.

Piano di check a posteriori su:

- *Variabili qualitative (deterministico e probabilistico)*
- *Variabili quantitative su consumi e spese (da modello).*

Il controllo e il trattamento dei dati

Il questionario elettronico è stato predisposto in modo da segnalare automaticamente:

- *le incongruenze (o incompatibilità) fra le risposte fornite dall'intervistato a più quesiti del questionario.*
- *i valori "fuori range" nei quesiti che rilevano dati numerici.*

…ma un piano di controllo e correzione a posteriori sui dati di indagine è stato necessario per:

- *Correggere errori di percorso del CATI (es: quesiti non dovuti in seguito a ritorni da regole di incompatibilità).*
- *Sanare incoerenze e fuori range non sanate in corso di intervista.*
- *Errori casuali dovuti a occasionali malfunzionamenti del CATI (rari e di scarsa entità).*

Il controllo e il trattamento dei dati

Il piano di check per variabili qualitative adottato per l'Indagine sui consumi energetici delle famiglie utilizza moduli deterministici e probabilistici

DETERMINISTICO - procedura tesa ad intervenire su errori di tipo sistematico che si basa su interventi di tipo if…then…

PROBABILITSTICO – moduli di coerenza complessiva tra più variabili attraverso una serie di regole di incompatibilità. In presenza di incoerenze, le variabili da correggere sono quelle che danno origine al minimo cambiamento nella struttura delle regole individuate. I valori che vengono successivamente imputati vengono scelti sulla base di donatori che hanno caratteristiche simili a quelli che subiscono l'imputazione. Il software utilizzato (CONCORD) è sviluppato e certificato dall'ISTAT.

Il controllo e il trattamento dei dati

Le variabili quantitative dell'indagine sono state sottoposte inizialmente ad un metodo di identificazione degli outlier attraverso l'approccio basato sul metodo non parametrico di Hidiroglou & Berthelot (1986).

Imputazioni attraverso modello di regressione log-lineare dove la variabile dipendente (spesa o consumo) è la trasformata logaritmica della variabile di partenza:

$$ln\,(y_i) = X_{i\_puliti} * \beta + \varepsilon$$

La stima assume una forma funzionale più vicina ad una gaussiana, il che assicura alcune proprietà fondamentali (Best Linear Unbiased) dello stimatore dei minimi quadrati (OLS).

Il controllo e il trattamento dei dati

Si deve però tenere conto della trasformazione logaritmica di partenza per attenuare l'effetto del retransformation bias (Duan (1983))

Vale difatti la seguente diseguaglianza:

$$E(yi) \neq e^{Xi\_da\ imputare\ *\ \beta}$$

E' stato applicato un correttore, smearing estimator, per ridurre questo tipo di distorsione basato sui residui della regressione log lineare che risulta robusto rispetto all'eteroschedasticità

La validazione con le altre fonti disponibili

- **ISTAT** - Censimento della popolazione e delle abitazioni 2011 (impianto fotovoltaico, impianto di riscaldamento abitazione, combustibile riscaldamento abitazione, combustibile riscaldamento acqua, impianto di condizionamento abitazione).

- **ISTAT** - Indagine sui consumi delle famiglie 2012 e 2013 (superficie abitazione, numero stanze, anno costruzione immobile, servizi abitazione, impianto di riscaldamento abitazione, spese per metano, spese per energia elettrica).

- **ISTAT** - Indagine sulle condizioni di vita (EU-SILC) 2012 e 2013 (spese per metano, spese per energia elettrica).

La validazione con le altre fonti disponibili

- **ISTAT - Indagini sugli aspetti della vita quotidiana 2013 (titolo di godimento dell'abitazione, numero stanze, servizi abitazione, impianto di riscaldamento abitazione, elettrodomestici).**
- **Ministero delle Sviluppo Economico - Bilancio Energetico Nazionale 2013.**
- **ENEA - Bilancio Energetico Regionale 2008.**
- **ISTAT – COEWEB data warehouse delle statistiche sul commercio estero.**
- **ENEA – indagine sui consumi energetici di biomasse nel settore residenziale in italia nel 1999.**
- **Indagini regionali su utilizzo di biomasse.**

Strategia di campionamento: disegno

Il disegno di campionamento adottato è stato realizzato dalla struttura dell'Istat, *Metodi per la progettazione di strategie campionarie innovative* è a uno stadio con stratificazione dei comuni all'interno della stessa regione per:

- ampiezza demografica e zona altimetrica che danno luogo a 7 tipologie comunali (Comune centro dell'area metropolitana; Comuni della periferia dell'area metropolitana; Comuni con più di 50.000 abitanti; Comuni di montagna interna, montagna esterna e collina interna con numero di abitanti tra 10mila e 50mila; Comuni di montagna interna, montagna esterna e collina interna con numero di abitanti fino a 10.000; Comuni di collina litoranea e pianura con numero di abitanti tra 10mila e 50mila; Comuni di collina litoranea e pianura con numero di abitanti fino a 10.000).

L'intero territorio nazionale è stato suddiviso in 118 strati.

All'interno di ogni strato, le famiglie campione sono state estratte casualmente dall'archivio delle famiglie abbonate alla rete di telefonia fissa (CONSODATA).

Oltre alle 20.000 famiglie che compongono il campione base, è stato predisposto un campione di **100000** famiglie di riserva, al fine di raggiungere la numerosità campionaria prefissata.

Strategia di campionamento: stima – *la procedura*

- Attribuzione del coefficiente di riporto all'universo che deriva direttamente dal disegno (pesi base) → inverso della probabilità di inclusione (per effetto del campionamento ogni unità osservata rappresenta una quota parte della popolazione che non fa parte del campione);
- Calcolo del correttore per mancata risposta *(aumento della rappresentatività del campione)*;
- Determinazione dei coefficienti di riporto all'universo finali. Gli istituti nazionali di statistica europei – e non solo – usano lo stimatore di ponderazione vincolata, **calibration estimator** (*Deville J.C., Särndal C.E., 1992; Särndal C.E., 2007*).

Strategia di campionamento: stima - *lo stimatore a ponderazione vincolata*

$$\hat{Y}_{cal} = \sum_{j \in s} w_j y_j$$

$$\min_{j \in s} \sum_{j \in s} G_j\left(w_j, a_j\right) \qquad \text{con il vincolo: } \sum_{j \in s} w_j x_j = X$$

- I pesi $\mathbf{w}_j$ devono essere tali da:
  - minimizzare la distanza dai pesi base $\mathbf{a}_j$;
  - riprodurre i totali noti $\boldsymbol{X}$ delle variabili ausiliarie.

Strategia di campionamento: stima - *La scelta delle informazioni ausiliarie*

- Lo stimatore a ponderazione vincolata garantisce le proprietà di **ottimalità** nel caso in cui si riescano a scegliere informazioni ausiliarie correlate con i principali parametri che l'indagine deve stimare.
- Nel caso delle indagini nelle quali si procede a selezione da liste telefoniche, il problema della **rappresentatività** del campione, a causa della *sottocopertura* delle liste telefoniche, è molto importante e, se non risolto, può avere ricadute negative sulla qualità delle stime prodotte.
- Nel caso dei consumi energetici, abbiamo usato, come variabile ausiliaria, il **reddito** complessivo delle famiglie partendo dall'**ipotesi** che, *a parità di altre condizioni*, *famiglie con redditi simili hanno la stessa tipologie di consumo energetico*.

Strategia di campionamento: stima - *La scelta delle informazioni ausiliarie –> le fonti*

1 - Banca dati reddituale (redditi 2011) del MEF

Contiene informazioni sui contribuenti che:

- hanno compilato il modello *UNICO Persone Fisiche*;
- hanno compilato il modello 730;
- sono presenti solo nei modelli 770.

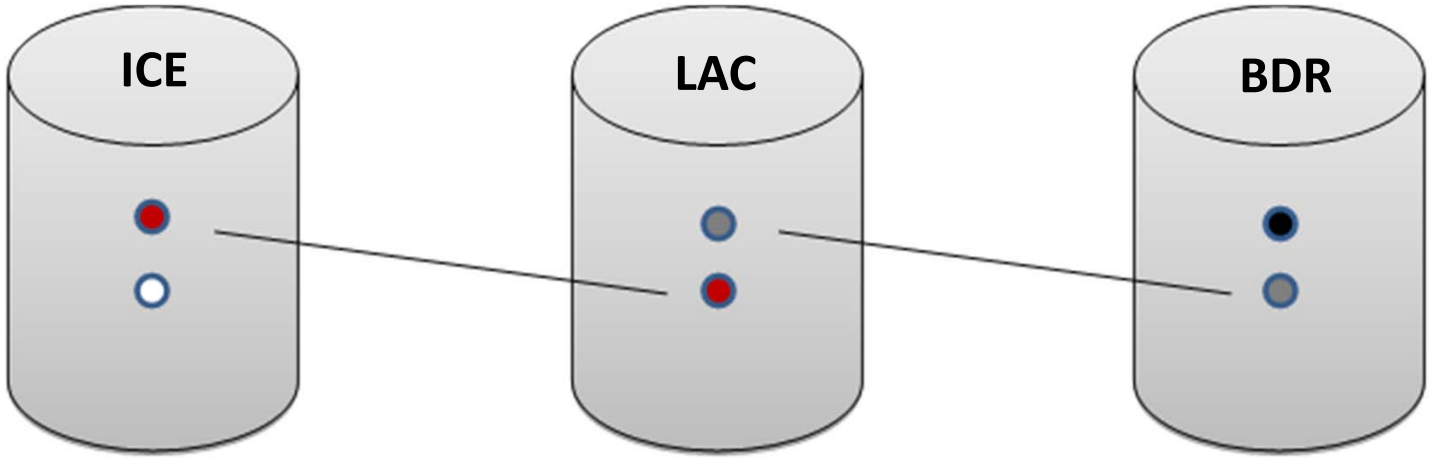
2 - Anagrafi comunali - LAC

Contiene informazioni su:

- popolazione residente;
- famiglie e convivenze;
- cittadini italiani e stranieri.

Strategia di campionamento: stima – *Il linkage tra le fonti*

La strategia prevede l'attribuzione del reddito familiare alle famiglie campione *(Record Linkage dei dati)*.



Chiave di aggancio **codice fiscale**

Chiave di aggancio **codice provincia || comune || famiglia**

Strategia di campionamento: stima - *La scelta delle informazioni ausiliarie –> le fonti*

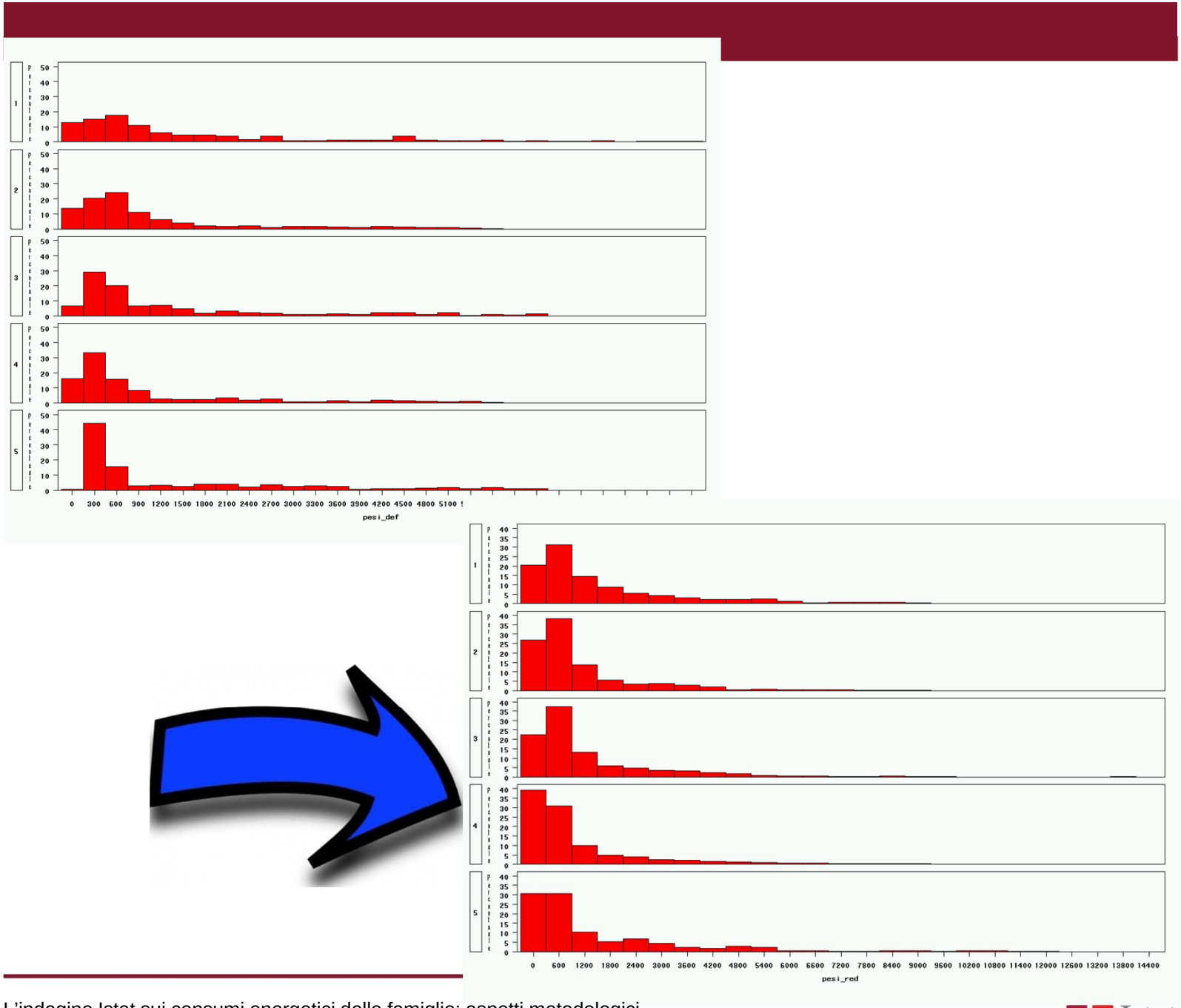
	FONTE	CHIAVE DI LINKAGE	FONTE	VARIABILI ACQUISITE	% di linkage
PRIMA FASE	ICE	TELEFONO	CONSODATA	**CODICE FISCALE** SUI RECORD DELL'INDAGINE CON PARTICOLARE ATTENZIONE AGLI ESITI=10 (INTERVISTE COMPLETE)	95%
SECONDA FASE	ICE	CODICE FISCALE	LAC	**CODICE FAMIGLIA** , **COMUNE** e **PROVINCIA**	91%
TERZA FASE	ICE	PROV‖COM‖CODICE_FAM	BDR	**REDDITO FAMILIARE** (SOMMA DEI REDDITI INDIVIDUALI)	91%

Strategia di campionamento: stima – *I correttori*

- Il reddito complessivo familiare è stato usato come variabile ausiliaria nel procedimento di calcolo dei coefficienti di riporto all'universo.
- Costruzione di un **fattore correttivo** che, moltiplicato per il peso base, riproduce la distribuzione del reddito della popolazione.
- Ogni famiglia italiana è collocata in una cella costruita dal prodotto cartesiano dei quinti di reddito e ripartizione geografica.
- All'interno di ogni casella è stata eseguita una vera e propria **riponderazione** partendo dall'ipotesi (forte, senza dubbio, ma molto plausibile) che ogni famiglia rispondente sia rappresentativa di un certo numero di famiglie all'interno della casella stessa.

La strategia di campionamento: la stimatore finale – *I vincoli usati nello stimatore a ponderazione vincolata*

- Al fine di aumentare l'efficienza delle stime e per garantire la coerenza delle stime prodotte con le informazioni ausiliarie note si procede alla definizione di vincoli di calibrazione sulla base della conoscenza di totali di popolazione noti da fonti esterne all'indagine.
  I totali utilizzati sono:
  1) popolazione residente per sesso e classe di età nelle cinque ripartizioni territoriali (Nord Est, Nord Ovest, Centro, Sud e Isole);
  2) popolazione residente per regione (incluse Trento e Bolzano);
  3) numero di famiglie residenti per regione;
  4) popolazione di 15 anni e oltre per condizione professionale e posizione nella professione (lavoratori alle dipendenze, autonomi, disoccupati, altri);
- I primi tre totali sono 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).


pesi_def
pesi_red

Grazie per l’attenzione!

Appendice – Le componenti del reddito complessivo familiare

Reddito Complessivo

È dato dalla somma dei singoli redditi indicati nei vari quadri:

- DOMINICALI
- AGRARI
- FABBRICATI
- da LAVORO DIPENDENTE
- da LAVORO AUTONOMO
- d'IMPRESA IN CONTABILITÀ ORDINARIA
- d'IMPRESA IN CONTABILITÀ SEMPLIFICATA
- d'IMPRESE CONSORZIATE
- da PARTECIPAZIONE (a)
- derivanti da PLUSVALENZE DI NATURA FINANZIARIA
- ALTRI REDDITI
- da ALLEVAMENTO
- TASSAZIONE SEPARATA (con opzione tassazione ordinarie) e pignoramento presso terzi

Riassume tutti i dati dichiarati negli altri quadri di questo modello (RN), utili per determinare l'imposta sui redditi delle persone fisiche (IRPEF) dovuta per l'anno d'imposta