ANALISI DELL'IMPATTO DELLA NUOVA STRATEGIA DI CAMPIONAMENTO DELL'INDAGINE ISTAT SUI CONSUMI DELLE FAMIGLIE

Claudia De Vitiis - Stefano Falorsi
Servizio Studi Metodologici

Riassunto

L'indagine sui consumi di famiglia ha subito un ridisegno a partire dal 1997. In questo documento viene descritta la strategia di campionamento della nuova indagine, mettendo in evidenza le modifiche introdotte rispetto alla strategia della vecchia indagine. Dopo aver illustrato i principali obiettivi conoscitivi dell'indagine, per mettere in luce gli aspetti maggiormente legati alla progettazione della nuova strategia, e dopo aver presentato brevemente la vecchia strategia, si descrive la nuova strategia evidenziando obiettivi e vincoli su cui è basata e approfondendo alcuni importanti aspetti metodologici che hanno caratterizzato il ridisegno. Viene riportata, infine, un'analisi dell'impatto delle due strategie sull'efficienza delle stime di spesa e sui livelli delle medesime stime.

1. Premessa

Il presente documento ha lo scopo di descrivere la strategia di campionamento della *nuova* indagine sui consumi di famiglia, mettendo in evidenza le modifiche introdotte rispetto alla strategia della *vecchia* indagine. L'esposizione è articolata come segue. Si illustrano dapprima i principali obiettivi conoscitivi dell'indagine, al fine di mettere in luce gli aspetti maggiormente legati alla progettazione della nuova strategia. Successivamente, dopo aver presentato brevemente la *vecchia* strategia, si descrive la *nuova* strategia evidenziando obiettivi e vincoli su cui è basata e approfondendo alcuni importanti aspetti metodologici che hanno caratterizzato il ridisegno. Viene riportata, infine, un'analisi dell'impatto delle due strategie sull'efficienza delle stime di spesa e sui livelli delle medesime stime.

2. Obiettivi conoscitivi dell'indagine

Oggetto dell'*Indagine sui consumi delle famiglie* sono la struttura e il livello dei consumi, secondo modalità di natura economica, sociale e territoriale delle famiglie e delle persone che le compongono. Inoltre l'indagine, essendo di tipo continuativo, consente di seguire e valutare l'evoluzione degli standard di vita familiari, in senso qualitativo e quantitativo. La *popolazione di interesse* è costituita dalle famiglie residenti in Italia, con

persone che le compongono. Inoltre l'indagine, essendo di tipo continuativo, consente di seguire e valutare l'evoluzione degli standard di vita familiari, in senso qualitativo e quantitativo. La *popolazione di interesse* è costituita dalle famiglie residenti in Italia, con esclusione quindi delle persone appartenenti in modo permanente alle convivenze e delle persone abitualmente residenti all'estero. L'*unità di rilevazione* è la *famiglia di fatto* intesa come insieme di persone coabitanti e legate da vincoli di matrimonio, parentela affinità adozione, tutela o affettivi. Sono considerate facenti parte della famiglia tutte le persone che, a qualsiasi titolo, convivono abitualmente con la famiglia (ISTAT, 1997).

Le stime fornite dall'indagine in oggetto sono espresse in termini di spesa media mensile per famiglia e per componente.

Gli ambiti territoriali rispetto ai quali l'indagine è chiamata a fornire stime campionarie sono:

- l'intero territorio nazionale;
- le cinque grandi ripartizioni geografiche;
- le ventuno regioni geografiche (comprese Bolzano e Trento).

Le stime hanno come riferimento temporale il trimestre o l'anno: le stime trimestrali sono richieste dalla Contabilità Nazionale a livello nazionale, mentre le stime annuali sono pubblicate dall'ISTAT con dettaglio regionale, ripartizionale e nazionale.

3. Disegno di campionamento della vecchia indagine

La strategia di campionamento della *vecchia* indagine non è studiata per rispondere agli obiettivi conoscitivi dell'indagine stessa. Come accadeva per la maggior parte delle rilevazioni campionarie ISTAT progettate fino all'inizio degli anni '80, il campione dell'indagine era ottenuto come sottocampione a partire dal campione dell'indagine sulle *forze di lavoro* (utilizzato fino ad aprile 1990).

Per la formazione del campione i comuni sono suddivisi in due gruppi:

- il primo include i comuni con più di 50.000 abitanti e i capoluoghi di provincia con meno di 50.000 abitanti (in tutto 150 comuni);
- il secondo è costituito dai rimanenti comuni.

Per il primo gruppo si adotta un *disegno a grappoli semplice stratificato*, in cui ciascun comune costituisce strato a se stante e viene pertanto indicato come Auto Rappresentativo (AR).

Per il secondo gruppo di comuni, indicati come comuni Non Auto Rappresentativi (NAR), si adotta un *disegno a due stadi con stratificazione delle unità di primo stadio*. I comuni vengono stratificati, all'interno di ciascuna regione geografica, secondo tre variabili: dimensione demografica, zona altimetrica e attività economica prevalente, operando di fatto un'aggregazione degli strati dell'indagine sulle *forze di lavoro*, per la quale la stratificazione, secondo le tre suddette variabili, veniva fatta all'interno delle provincie.

Da ogni strato NAR si selezionano, mediante scelta ragionata, tre comuni tra quelli appartenenti al campione di comuni delle *forze di lavoro*, per un totale di 405 comuni.

Da ogni comune campione si estrae una predeterminata frazione di famiglie anagrafiche.

Ogni trimestre partecipano all'indagine 9.000 famiglie, per un totale di 36.000 famiglie annue.

La distribuzione del campione nel tempo è di seguito schematizzata:

- ogni comune AR è coinvolto nell'indagine tutti i mesi;
- in ogni strato NAR, ciascuno dei tre comuni campione di partecipa all'indagine quattro volte nell'anno, a distanza di tre mesi l'una dall'altra, secondo lo schema seguente:

Prospetto 1. Schema di partecipazione dei comuni campione NAR nei mesi dell'anno

comuni	**mese di rilevazione**			
1	gennaio	aprile	luglio	ottobre
2	febbraio	maggio	agosto	novembre
3	marzo	giugno	settembre	dicembre

L'estrazione delle famiglie campione viene effettuata una volta l'anno mediante selezione sistematica dalle anagrafi comunali; le famiglie estratte sono ripartite nell'anno come segue:

- il campione di famiglie di ciascun comune AR è ripartito in 12 campioni mensili, mentre

L'estrazione delle famiglie campione viene effettuata una volta l'anno mediante selezione sistematica dalle anagrafi comunali; le famiglie estratte sono ripartite nell'anno come segue:

- il campione di famiglie di ciascun comune AR è ripartito in 12 campioni mensili, mentre il campione di famiglie di ciascun comune NAR è suddiviso in 4 campioni mensili;
- ciascun campione mensile di famiglie viene suddiviso in tre sottoinsiemi: il primo sottoinsieme viene rilevato la prima decade del mese, il secondo nella seconda decade ed il terzo nell'ultima decade;
- l'operazione di osservazione consiste di due fasi: la famiglia annota nel *libretto degli acquisti* le spese giornaliere per generi alimentari e per alcuni generi non alimentari di tipo non durevole; alla fine del mese il rilevatore raccoglie le informazioni relative alle altre variabili di spesa che possono avere diversi periodi di riferimento, quali il mese, il trimestre o l'anno.

Per quanto riguarda la procedura di stima della vecchia indagine si rimanda al paragrafo 7 in cui verranno esaminati gli stimatori utilizzati per la vecchia e per la nuova indagine.

4. La revisione della strategia di campionamento

La principale finalità alla base della revisione della strategia di campionamento dell'indagine sui consumi è stata sia quella di studiare una strategia maggiormente rispondente agli obiettivi conoscitivi dell'indagine stessa, sia quella di adottare anche per tale indagine le innovazioni metodologiche – volte al miglioramento della qualità delle stime - già introdotte nella ristrutturazione o nella progettazione delle altre importanti indagini sulla popolazione, quali l'indagine sulle *forze di lavoro* e l'indagine *multiscopo* sulle famiglie. Ovviamente il ridisegno dell'indagine è stato anche guidato dalla necessità di rispettare vincoli di carattere operativo e di costo derivanti dalle operazioni di rilevazione sul campo. Le principali modifiche del disegno d'indagine derivanti dai vincoli operativi e di costo sono le seguenti:

- la numerosità campionaria teorica subisce una riduzione sia in termini di comuni che in termini di famiglie,
- i capoluoghi di provincia sono tutti coinvolti nella rilevazione ogni mese (per ragioni legate alla rilevazione mensile dei prezzi);
- il periodo di rilevazione passa da dieci a sette giorni;
- il campione mensile di ogni comune viene ripartito in due sottoinsiemi da rilevare in due settimane campione del mese;
- le due settimane campione di ciascun comune devono essere separate tra loro da almeno sette giorni e contenute interamente nel mese;
- ogni mese i comuni campione di ogni regione svolgono la rilevazione nelle stesse due settimane campione.

Per quanto riguarda, invece, gli aspetti propriamente metodologici del ridisegno, dal momento che l'indagine utilizza ancora l'intervista diretta, si è mantenuto lo schema a due stadi comuni-famiglie e la suddivisione dei comuni nei due sottogruppi AR e NAR. Le principali modifiche che sono state introdotte nella strategia sono:

- la distribuzione del campione tra le regioni – in termini di comuni e famiglie - viene studiata sulla base di criteri di efficienza delle stime relative ai diversi domini territoriali di studio;
- la stratificazione dei comuni NAR viene effettuata nell'ambito di ciascuna regione geografica utilizzando la sola dimensione demografica dei comuni, sotto il vincolo che gli strati abbiano approssimativamente la medesima ampiezza;
- la *soglia* di popolazione che definisce i comuni AR viene stabilita, basandosi su criteri operativi e di riduzione degli errori campionari, in modo differente da regione a regione;
- la procedura di stima è quella generalmente utilizzata in tutte le indagini ISTAT ed è basata su uno stimatore di *ponderazione vincolata* (Falorsi, Rinaldelli, 1998).

Le fasi che hanno condotto alla definizione del nuovo disegno di campionamento vengono illustrate nel paragrafo 5, mentre nel paragrafo 6 vengono trattati in modo più approfondito alcuni aspetti relativi alla definizione della numerosità campionaria e alla stratificazione dei comuni.

5. Fasi di definizione del nuovo disegno

Specificazione delle unità campionarie e dei domini di stima. $(l=1,K,K,L)$ dei domini di

5. Fasi di definizione del nuovo disegno

Con riferimento alla generica regione r, si denoti con: l $(l=1,K\ K,L)$, l'indice di strato di comuni; c $(c=1,K\ K,C)$, l'indice di comune; N_l, il numero di famiglie residenti nello strato l; N_{lc} il numero di famiglie residenti nel comune c dello strato l; P_l, il numero di individui residenti nello strato l; P_{lc} il numero di individui residenti nel comune c dello strato l; m, il numero di comuni campione in ogni strato (m=1 per gli strati AR ed m=3 per gli strati NAR).

I criteri che sono stati seguiti per la definizione del disegno campionario sono:

- autoponderazione del campione al livello di regione;
- definizione di un numero minimo di famiglie da intervistare per comune;
- stratificazione dei comuni sulla base dell'ampiezza demografica;
- formazione di strati di comuni di ampiezza approssimativamente costante in termini di popolazione residente.

Per la definizione del campione di comuni e di famiglie relativo ad un trimestre sono state dapprima effettuate le seguenti scelte:

(a) definizione del numero complessivo di famiglie campione a livello nazionale: 24.000 famiglie annue ripartite in 6.000 famiglie al trimestre;

(b) definizione del numero n_r di famiglie campione per ciascuna regione;

(c) scelta del numero minimo di famiglie, $\bar{n}_r$, da intervistare in ciascun comune campione;

Dalla scelta di $\bar{n}_r$ e n_r (effettuata secondo criteri che saranno approfonditi nel punto C. del paragrafo 6) dipende la suddivisione dei comuni in AR e NAR e la formazione degli strati attraverso i seguenti passi:

(1) calcolo della frazione di campionamento regionale $f_r = n_r / N_r$, essendo N_r il numero di famiglie residenti nella regione r ;

(2) determinazione del valore della *soglia* λ_r, mediante la relazione

$$\lambda_r = \frac{\bar{n}_r \delta_r}{f_r},$$

in cui δ_r è il numero medio di componenti per famiglia a livello regionale; risulta evidente da tale espressione che la soglia per la definizione dei comuni AR cresce al crescere di $\bar{n}_r$;

(3) suddivisione dei comuni in AR e NAR sulla base della soglia λ_r (i comuni capoluoghi di provincia sono comunque definiti AR anche se non superano la soglia);

(4) ordinamento decrescente dei comuni NAR all'interno di ogni regione in funzione della loro dimensione demografica;

(5) suddivisione dei comuni NAR in strati la cui dimensione è approssimativamente uguale al prodotto $m \times \lambda_r$;

(6) selezione di m=3 comuni campione da ciascuno strato l (l=1,…,L) con probabilità proporzionale all'ampiezza; per il generico comune c tale probabilità è espressa dalla formula

$$z_{lc} = P_{lc} / P_l ;$$

(7) definizione del numero n_{lc} di famiglie da intervistare in ogni comune; dalla condizione di autoponderazione a livello regionale

$$\frac{m P_{lc}}{P_l} \frac{n_{lc}}{N_{lc}} = f_r ,$$

in cui il primo membro rappresenta la probabilità d'inclusione delle famiglie del comune c dello strato l (essendo le due frazioni rispettivamente la probabilità d'inclusione di primo e di secondo stadio), si ottiene

$$n_{lc} = \frac{f_r}{m} \frac{P_l N_{lc}}{P_{lc}} .$$

Una volta estratto il campione di comuni, lo schema di partecipazione dei comuni durante l'anno è il seguente:

- ogni comune AR è coinvolto nell'indagine tutti i mesi;
- [illegible]

Una volta estratto il campione di comuni, lo schema di partecipazione dei comuni durante l'anno è il seguente:

- ogni comune AR è coinvolto nell'indagine tutti i mesi;
- ciascuno dei tre comuni campione di ogni strato NAR partecipa all'indagine quattro mesi nell'anno a distanza di tre mesi secondo lo stesso schema della vecchia indagine.

L'estrazione delle famiglie campione viene effettuata una volta l'anno mediante selezione sistematica dalle anagrafi comunali; per ciascun comune campione vengono selezionate quindi complessivamente (cfr. punto 7) $4n_{lc}$ famiglie che vengono ripartite nell'anno come segue:

- il campione di famiglie di ciascun comune AR viene suddiviso in 12 campioni mensili;
- il campione di famiglie di ciascun comune NAR viene suddiviso in 4 campioni mensili.
- ciascun campione mensile viene suddiviso in due gruppi, uno per ciascuna settimana campione del mese.

La selezione delle due settimane di rilevazione per ciascun mese dell'anno avviene secondo le seguenti modalità: poiché ogni mese, per ragioni di tipo organizzativo, tutti i comuni campione di una data regione devono svolgere l'indagine nelle stesse due settimane, tali settimane vengono selezionate indipendentemente da regione a regione e tra un mese e l'altro, mediante un'estrazione casuale che garantisce la sostanziale equiprobabilità di inclusione dei giorni del mese.

Nel prospetto seguente vengono riportate le numerosità campionarie in termini di comuni e famiglie.

Prospetto 2. Numerosità del campione di comuni e famiglie

	Campione					
	Comuni			**Famiglie**		
	AR	NAR	TOTALE	AR	NAR	TOTALE
Mese	107	127	234	720	1266	1986
Trimestre	107	381	488	2160	3798	5958
Anno	107	381	488	8640	15192	23832

6. Approfondimenti metodologici

6.1 Scelta delle variabili di stratificazione

In generale, l'obbiettivo della stratificazione è quello di formare gruppi (o strati) di unità statistiche caratterizzati da:

- massima omogeneità interna agli strati rispetto alle variabili di interesse;
- massima differenza di comportamento delle variabili fra i diversi strati.

Il raggiungimento di tale obbiettivo si traduce in termini statistici in guadagno nella precisione delle stime, ossia in una riduzione dell'errore campionario a parità di numerosità campionaria.

Si ritiene inoltre utile sottolineare i seguenti aspetti:

- la stratificazione delle UP in un campionamento a due stadi conduce ad aumenti comunque modesti dell'efficienza delle stime (Kish, 1965);
- la stratificazione viene adottata essenzialmente per garantire la dimensione del campione in definiti domini di studio;
- per ciascuna variabile esiste una stratificazione di tipo ottimale; l'individuazione di tale ottimo implica la conoscenza delle variabili oggetto d'indagine su tutte le unità della popolazione. Per ogni variabile oggetto d'indagine, l'attuazione della stratificazione ottima su una variabile ausiliaria ad essa correlata comporta comunque una perdita di efficienza. Tale perdita è tanto maggiore quanto più debole è la correlazione suddetta;
- un ulteriore elemento che riduce l'importanza della stratificazione è relativo al fatto che essa viene attuata unicamente sui comuni appartenenti all'insieme NAR, mentre le stime campionarie sono costruite considerando sia l'insieme AR che quello NAR. Per tale ragione, anche un notevole guadagno di efficienza sulla parte NAR ha un effetto ridotto sull'efficienza della stima complessiva degli insiemi AR e NAR
- le ricerche condotte in Italia in tema di stratificazione pur non essendo numerose sono tuttavia ricche di risultati significativi. Tali ricerche, alcune delle quali condotte in ambito ISTAT (Zannella, 1991), hanno guidato verso la scelta di utilizzare nelle indagini ISTAT sulle famiglie la sola dimensione demografica dei comuni per la stratificazione dei medesimi, in luogo di altre variabili (ad esempio, altitudine, e attività economica

tuttavia ricche di risultati significativi. Tali ricerche, alcune delle quali condotte in ambito ISTAT (Zannella, 1991), hanno guidato verso la scelta di utilizzare nelle indagini ISTAT sulle famiglie la sola dimensione demografica dei comuni per la stratificazione dei medesimi, in luogo di altre variabili (ad esempio, altitudine, e attività economica prevalente) utilizzate precedentemente.

Per le ragioni ora illustrate non è vantaggioso dedicare molte risorse all'individuazione di una stratificazione *ottima*. Risulta più conveniente, quindi, adottare una soluzione metodologica che sia operativamente facile, fattibile e che rispetti condizioni di ragionevolezza. In questa ottica, per l'indagine in esame, si è ritenuto opportuno attuare la stratificazione dei comuni in base alla sola dimensione demografica.

6.2. Formazione degli strati

La formazione degli strati qui suggerita rientra nella logica della determinazione ottimale dei confini suggerita da Mahalanobis (1952) e da Hansen, Hurwitz e Madow (1953) la quale consiste nel delimitare gli strati in modo che la dimensione sia all'incirca costante.

Nel prosieguo si ritiene utile approfondire il significato metodologico di tale scelta, facendo riferimento al disegno di campionamento della nuova indagine e considerando la parte dei comuni NAR. A tale scopo, riprendendo la notazione del paragrafo 5, si introducono, con riferimento alla generica regione, i seguenti indici: g, indice di classe di strati, essendo ciascuna classe g, $(g = 1,K\ K,G)$, un raggruppamento di L_g $(l = 1,K\ K,L_g)$ strati; j indice di famiglia nel comune c dello strato l della classe g. Con riferimento al generico strato l della classe g, si indichi, poi, con: M_{gl}, numero di comuni universo; m_{gl}, numero di comuni campione; N_{glc} numero di famiglie universo nel comune c; n_{glc} numero di famiglie campione nel comune c; Y_{glcj}, il valore della variabile y oggetto di indagine osservato sulla famiglia j del comune c.

Il parametro oggetto di stima è il totale della generica variabile y oggetto di indagine, espresso dalla relazione seguente

$$Y = \sum_{g=1}^{G} Y_g = \sum_{g=1}^{G}\sum_{l=1}^{L_g} Y_{gl} = \sum_{g=1}^{G}\sum_{l=1}^{L_g}\sum_{c=1}^{\bar{m}_g} Y_{glc} = \sum_{g=1}^{G}\sum_{l=1}^{L_g}\sum_{c=1}^{\bar{m}_g}\sum_{j=1}^{N_{glc}} Y_{glc} \quad .$$

Una stima corretta del parametro Y è data da

$$Y = \sum_{g=1}^{G} \frac{1}{m} \sum_{l=1}^{L_g}\sum_{c=1}^{m} \frac{Y_{glc}}{N_{glc}}$$

in cui

$$Y_{glc} = \sum_{j=1}^{n_{glc}} \frac{N_{glc}}{n_{glc}} Y_{glcj}$$

è una stima corretta del totale Y_{glc}. E' utile ai nostri fini esprimere la varianza campionaria della stima Y mediante la seguente formula

$$Var(Y) = \sum_{g=1}^{G} \frac{1}{m} \sum_{l=1}^{L_g} Y_{gl}^2\, \varepsilon_{gl}^2 \quad ,$$

dove

$$\varepsilon_{gl}^2 = \frac{1}{M_{gl}\, Y_{gl}^2} \sum_{c=1}^{M_{gl}} N_{glc} \left(\frac{Y_{glc}}{N_{glc}} - Y_{gl} \right)^2$$

rappresenta la varianza relativa tra i comuni all'interno dello strato l della classe g.

Nell'ipotesi che la varianza relativa ε_{gl}^2 rimanga la stessa in tutti gli L_g strati appartenenti a ciascuna classe g $(g = 1,K\ K,G)$, l'insieme dei valori di Y_{gl} $(l = 1,K\ K,L_g)$ che minimizzano la varianza della stima Y si ottiene come soluzione di un problema di minimo vincolato in cui la funzione obiettivo è data da

$$\min \left\{ \sum_{l=1}^{L_g} Y_{gl}^2\, \varepsilon_{gl}^2 \right\} \qquad (g = 1,K\ K,G)$$

sotto il vincolo

$$[\quad\quad] \qquad (g = 1,K\ K,G)$$

sotto il vincolo

$$\sum_{l=1}^{L_g} Y_{gl} = Y_g \qquad (g = 1,K\ K,G)\ .$$

Al fine di ottenere i valori di $Y_{gl}\ (l = 1,K\ K,L_g)$ soluzione del problema di minimo vincolato si definisce la seguente funzione di Lagrange

$$F = \sum_{l=1}^{L_g} Y_{gl}^2\ \varepsilon_{gl}^2 + \lambda\left(\sum_{l=1}^{L_g} Y_{gl} - Y_g\right) \qquad (g = 1,K\ K,G)\ .$$

Si risolve, quindi, il sistema omogeneo di $(L_g + 1)$ equazioni nelle $(L_g + 1)$ incognite Y_{gl} ($(l = 1,K\ K,L_g)$, l)

$$\frac{\partial F}{\partial Y_{gl}} = 2\ Y_{gl}\ \varepsilon_{gl}^2 + \lambda = 0 \qquad \text{per } (g = 1,K\ K,G) \text{ e } (l = 1,K\ K,L_g)$$

$$\frac{\partial F}{\partial \lambda} = Y_{gl} - Y_g = 0 \qquad \text{per } (g = 1,K\ K,G) \text{ e } (l = 1,K\ K,L_g)$$

Risolvendo rispetto a Y_{gl} si ottiene mediante semplici passaggi la soluzione cercata

$$Y_{gl} = \frac{Y_g}{\sum_{l=1}^{L_g} \frac{1}{\varepsilon_{gl}^2}}\ \frac{1}{\varepsilon_{gl}^2} \qquad \text{per } (g = 1,K\ K,G) \text{ e } (l = 1,K\ K,L_g)\ .$$

poiché il primo termine a secondo membro è espresso dal rapporto tra due quantità costanti, Y_{gl} e $\sum_{l=1}^{L_g} \frac{1}{\varepsilon_{gl}^2}$, è possibile scrivere la precedente espressione come

$$Y_{gl} = K\ \frac{1}{\varepsilon_{gl}^2} \qquad \text{per } (g = 1,K\ K,G) \text{ e } (l = 1,K\ K,L_g)$$

in cui

$$K = \frac{Y_g}{\sum_{l=1}^{L_g} \frac{1}{\varepsilon_{gl}^2}} \qquad \text{per } (g = 1,K\ K,G) \text{ e } (l = 1,K\ K,L_g)\ .$$

Essendo, quindi, per ipotesi la varianza relativa degli L_g strati della classe g $(g = 1,K\ K,G)$, ε_{gl}^2, una quantità costante, $\overline{\varepsilon}_g^2$, la scelta ottimale per la formazione degli strati è quella di formare strati di ampiezza approssimativamente costante, la cui ampiezza sia pari a

$$Y_{gl} = \overline{K}\ \frac{1}{\overline{\varepsilon}_g^2} \qquad \text{per } (g = 1,K\ K,G) \text{ e } (l = 1,K\ K,L_g)$$

in cui

$$\overline{K} = \frac{Y_g}{L_g\ \frac{1}{\overline{\varepsilon}_g^2}} \qquad \text{per } (g = 1,K\ K,G)\ .$$

6.3. Definizione della numerosità campionaria

Per un indagine ad obbiettivi plurimi come è quella in esame, è poco realistico pensare di potere disegnare una strategia campionaria che assicuri prefissati livelli di precisione a tutte le stime prodotte. La questione è complicata dal fatto che l'indagine ha la finalità di determinare stime per livelli territoriali e temporali differenti il che comporta in genere soluzioni di tipo ottimale diverse e contrastanti tra loro.

Si è deciso di adottare un ottica mista basata sia su criteri di costo ed organizzativi sia su una valutazione degli errori campionari delle principali stime a livello

genere soluzioni di tipo ottimale diverse e contrastanti tra loro.

Si è deciso di adottare un ottica mista basata sia su criteri di costo ed organizzativi sia su una valutazione degli errori campionari delle principali stime a livello nazionale e regionale. I criteri seguiti possono essere sintetizzati nei seguenti punti:

- la dimensione complessiva del campione annuo di famiglie a livello nazionale, in base a criteri di costo e operativi, non poteva superare una dimensione di 24.000;
- il numero di comuni campione coinvolti in ciascun trimestre non poteva essere superiore a 500; si è scelto tuttavia, per ragioni di efficienza campionaria, di diminuire il numero di comuni campione in misura minore di quanto fatto per le famiglie;
- l'allocazione del campione di famiglie e comuni tra le varie regioni è stata ottenuta adottando un criterio di compromesso che garantisse l'affidabilità sia delle stime regionali sia delle stime nazionali (Sing. et al. (1994)).

In particolare, si è pervenuti alla determinazione delle dimensioni campionarie a livello regionale nel seguente modo:

(A) si sono considerate tre allocazioni alternative:

- un'allocazione *uguale*, che conduce a stime che presentano approssimativamente il medesimo errore a livello regionale. Tale allocazione è poco efficiente per le stime nazionali;
- un'allocazione *proporzionale* alla dimensione demografica delle regioni stesse. Questa allocazione conduce a stime efficienti a livello nazionale, ma presenta l'inconveniente che le stime relative alle regioni piccole, come ad esempio la Val d'Aosta ed il Molise, sono affette da errori di campionamento molto elevati;
- un'allocazione di *compromesso* tra le due precedentemente illustrate, che, da una parte garantisce che l'efficienza delle stime nazionali non sia molto distante da quella ottima (ottenuta con l'allocazione proporzionale), dall'altra parte assicura un minimo di affidabilità alle stime delle regioni più piccole;

(B) si sono considerate diverse modalità di suddivisione del campione di comuni tra parte AR e parte NAR, considerando valori alternativi del numero minimo, $_r\bar{n}$, di interviste mensili da effettuare in ciascun comune campione, oltre alla possibilità di differenziare tale valore a seconda della dimensione della regione.

Con riferimento alla scelta del numero minimo, si fa presente che da essa dipende la numerosità del campione di comuni; esiste, infatti, (cfr. paragrafo 5) una relazione diretta tra $_r\bar{n}$ e la dimensione degli strati, inversamente legata a sua volta al numero di comuni campione. Di conseguenza, il numero minimo non può essere fissato solamente in base a criteri di efficienza delle stime, ma è necessario tenere anche conto di questioni di costo.

E' stato quindi considerato un insieme di disegni di campionamento ottenuti combinando le diverse alternative dei punti (a) e (b); per ognuno di tali disegni sono stati calcolati gli errori di campionamento attesi delle stime delle medie relative ai principali capitoli di spesa secondo la metodologia di seguito sinteticamente descritta (per una trattazione più dettagliata si veda Falorsi e Falorsi, 1996).

L'errore di campionamento atteso della stima della spesa media mensile nel trimestre per famiglia $_r\bar{Y}$ della regione r è definito da

$$\varepsilon(_r\bar{Y}) = \sqrt{\frac{_rS^2 \; _r\text{deff}}{_rn \; _r\bar{Y}^2}} = \sqrt{\frac{\text{Var}(_r\bar{Y})}{_r\bar{Y}^2}}$$

in cui: $\frac{_rS^2}{_rn}$ denota la varianza della stima $_r\bar{Y}$ ottenibile con un (ipotetico) campione casuale semplice di dimensione $_r$n, pari al numero complessivo di famiglie da includere nel campione e $_r$deff, noto con il termine effetto del disegno di campionamento, è un fattore moltiplicativo, generalmente superiore ad 1, che misura (relativamente alla stima $_r\bar{Y}$) l'efficienza del campione complesso rispetto a quella di un campione casuale semplice di pari numerosità (Kish, 1965). Una formula approssimata di definizione del deff, che tenga conto della ripartizione del campione tra la parte AR e quella NAR, è definita da

$$\text{deff} = \frac{_rn}{_rN^2}\left\{\left[\frac{_rN_{AR}^2}{_rn_{AR}} \; _r\text{deff}_{AR}\right] + \left[\frac{_rN_{NAR}^2}{_rn_{NAR}}\left(1 + \; _r\text{rho}_{NAR}\left(_r\bar{n} - 1\right)\right)\right]\right\},$$

in cui: $_rN_{AR}$ e $_rN_{NAR}$ indicano il totale delle famiglie nella parte AR e NAR; $_rn_{AR}$ ed $_rn_{NAR}$

$$ {}_rN^2 \left[\left[{}_rn_{AR} \right] \left[{}_rn_{NAR} \right] \right] , $$

in cui: ${}_rN_{AR}$ e ${}_rN_{NAR}$ indicano il totale delle famiglie nella parte AR e NAR; ${}_rn_{AR}$ ed ${}_rn_{NAR}$ denotano il numero di famiglie campione in AR e NAR; ${}_rdeff_{AR}$ è un fattore moltiplicativo, inferiore ad 1, che misura l'efficacia della stratificazione della parte AR; ${}_rrho_{NAR}$ indica il coefficiente di correlazione intraclasse in NAR.

L'errore relativo della stima $\overline{Y}$ a livello nazionale è definito da

$$\varepsilon(\overline{Y}) = \sqrt{\frac{\mathrm{Var}(\overline{Y})}{\overline{Y}^2}}$$

in cui

$${}_r\overline{Y} = \sum_{r=1}^{21} \frac{{}_rN}{N} {}_r\overline{Y} \quad ; \quad \mathrm{Var}(\overline{Y}) = \sum_{r=1}^{21} \left(\frac{{}_rN}{N} \right)^2 \mathrm{Var}({}_r\overline{Y}$$

dove N indica il totale delle famiglie a livello nazionale.

Per calcolare il valore numerico delle espressioni ora indicate si è proceduto nel modo seguente:

le quantità ${}_rdeff_{AR}$, e ${}_rrho_{NAR}$ sono state stimate sui dati dell'indagine del 1992;

le quantità rN ed N sono quelle desumibili dalle statistiche demografiche dell'ISTAT;

le quantità ${}_rN_{AR}$, ${}_rN_{NAR}$, ${}_rn$, ${}_rn_{AR}$, ${}_rn_{NAR}$, ${}_r\overline{n}$ sono quelle relative allo specifico disegno di campionamento adottato.

Il piano di campionamento prescelto è quello che prevede un'allocazione di *compromesso* ed un numero minimo ${}_r\overline{n}$ fissato pari a 10 per le regioni più piccole (Valle d'Aosta, Molise, Bolzano, Trento, Basilicata, Umbria, Friuli, Abruzzi Marche, Sardegna, Liguria e Calabria) e pari a 12 nelle rimanenti; tale scelta garantisce da un lato stime affidabili sia al livello di tutte le regioni e ripartizioni geografiche che al livello dell'intero territorio nazionale, dall'altro consente di mantenere il numero totale di comuni campione al di sotto dei 500.

7. Procedura di stima

In generale, per ottenere la stima di un totale devono essere eseguite tre operazioni: determinare il coefficiente di riporto (o peso) da attribuire a ciascuna unità inclusa nel campione, moltiplicare il valore relativo ad una data variabile oggetto di indagine, rilevata sulla generica unità inclusa nel campione, per il peso attribuito alla medesima unità ed effettuare la somma dei prodotti così ottenuti. Con riferimento alle due indagini messe a confronto, verranno descritte le due procedure che conducono alla definizione dei coefficienti di riporto.

La *vecchia* indagine adottava un metodo di stima in cui il peso di ogni famiglia campione della regione r, dello strato l e di ampiezza a, è ottenuto dal prodotto di due fattori:

- il primo fattore è dato dal rapporto, per ciascuno strato l della regione r, tra il totale noto della popolazione residente, al netto dei membri permanenti delle convivenze e il numero di individui campione

$${}_rc_l = \frac{{}_rP_l}{{}_rp_l} \quad ;$$

- il secondo fattore è dato dal rapporto, per ciascuna regione r e con riferimento ad ogni ampiezza familiare a, (a=1,...,6 e più) tra il totale delle famiglie - ricostruito sulla base dei dati censuari - e la corrispondente stima ottenuta ponderando i dati campionari sulla base del primo fattore; in simboli

$$w_{la} = {}_rc_l \frac{{}_rF_a}{{}_rF_a} \quad , \text{ essendo } \quad {}_rF_a = \sum_l {}_rc_l \, {}_rf_{la} \quad .$$

Il generico totale ${}_rF_a$ nella regione r viene ottenuto a partire dall'anno 1994 sulla base dei dati censuari. In particolare si opera nel seguente modo: con riferimento ai due domini AR e NAR di ciascuna regione geografica si calcolano i totali della popolazione residente al netto dei membri permanenti delle convivenze, indicati rispettivamente con $P_{\quad}$ $P_{\quad}$ si dividono

Il generico totale $_rF_{ar}$ nella regione r viene ottenuto a partire dall'anno 1994 sulla base dei dati censuari. In particolare si opera nel seguente modo: con riferimento ai due domini AR e NAR di ciascuna regione geografica si calcolano i totali della popolazione residente al netto dei membri permanenti delle convivenze, indicati rispettivamente con $_rP_{AR}$, $_rP_{NAR}$, si dividono poi tali totali per il numero medio di componenti per famiglia risultante al censimento, ottenendo in tal modo una stima del numero di famiglie del dominio AR, $_rF_{AR}$, e quello del dominio NAR, $_rF_{NAR}$, sommando poi tali totali si ottiene una stima del numero totale di famiglie della regione r, $_rF$, moltiplicando, infine, tale totale per la frequenza relativa di famiglie di ampiezza a, risultante dal censimento, si ottiene la stima cercata. Si ricorda che fino all'anno 1993 invece dei dati censuari venivano utilizzate le stime provenenti dall'indagine sulle forze di lavoro. L'utilizzo di una distribuzione esterna per ampiezza familiare di fonte censuaria può introdurre delle distorsioni nelle stime tanto maggiori quanto più la distribuzione reale si discosta da quella censuaria.

La procedura di stima utilizzata per la *nuova* indagine prevede che il peso da attribuire a ciascuna famiglia j si ottenga mediante una successione di passi.

1. viene calcolato un peso iniziale, d_j, definito *peso diretto*, e ottenuto come reciproco della probabilità d'inclusione delle unità nel campione;
2. viene calcolato un fattore correttivo c_j del peso diretto allo scopo di correggere la distorsione causata dalla mancata risposta totale e di rispettare la condizione di uguaglianza tra un vettore di parametri noti della popolazione, $\underline{X}$, e le corrispondenti stime campionarie, $\underline{X}_w$, ottenute con i pesi finali;
3. viene determinato infine il peso *finale* w_j espresso come prodotto del peso diretto per il fattore correttivo.

Il fattore correttivo c_j deriva dalla risoluzione di un problema di minimo vincolato. La funzione obbiettivo è data da

$$\min\left\{\sum_{j\in s} G_j(w_j;d_j)\right\},$$

in cui G(.;.) è un'opportuna funzione di distanza tra l'insieme dei pesi finali e l'insieme dei pesi diretti; i vincoli del problema sono espressi dal sistema di K equazioni

$$\underline{X}_w = \underline{X} \quad ,$$

I totali noti di riferimento utilizzati per la *nuova indagine*, definiti a livello di ciascuna ripartizione geografica, sono:

a) la distribuzione della popolazione della ripartizione per sesso e classi di età (0-14, 15-29, 30-59, 60 e più);
b) il totale della popolazione di ciascuna regione della ripartizione;
c) il totale delle famiglie anagrafiche per ciascuna regione della ripartizione.

Pertanto le variabili ausiliarie coinvolte nel procedimento di stima sono per ogni famiglia:

a) la distribuzione dei componenti della famiglia per sesso e classe di età (0-14, 15-29, 30-59, 60 e più);
b) il totale dei componenti;
c) la regione di residenza.

Le variabili ausiliarie che entrano così in gioco identificano in sostanza la tipologia familiare per sesso ed età rispetto alla quale, quindi, il campione viene post-stratificato.

La funzione di distanza utilizzata è la funzione *logit* che assicura la positività dei coefficienti di riporto e consente di limitare il campo di variazione dei pesi finali. Lo stimatore utilizzato appartiene alla famiglia degli stimatori di *ponderazione vincolata*, di cui è nota la tendenza asintotica allo stimatore di *regressione generalizzata*, che si ottiene nel caso in cui la funzione di distanza sia la funzione di distanza euclidea

$$G(w_{js};d_j) = \frac{(d_j - w_{js})^2}{d_j} \quad .$$

Tale stimatore ha la seguente espressione

$$Y_{reg} = Y + \underline{\beta}'(\underline{X} - \underline{X})$$

in cui Y e $\underline{X}$ sono gli stimatori diretti rispettivamente del totale di spesa Y e del vettore dei totali noti $\underline{X}$, mentre $\underline{\beta}$ è il vettore stimato dei coefficienti di regressione del sottostante modello di regressione lineare che lega la generica variabile di interesse y all'insieme delle

in cui $\hat{Y}$ e $\hat{\underline{X}}$ sono gli stimatori diretti rispettivamente del totale di spesa Y e del vettore dei totali noti $\underline{X}$, mentre $\hat{\underline{\beta}}$ è il vettore stimato dei coefficienti di regressione del sottostante modello di regressione lineare che lega la generica variabile di interesse y all'insieme delle variabili ausiliarie (tipologia familiare).

8. Lo stimatore della spesa mensile per consumi

Si consideri la variabile y che rappresenta la spesa effettuata da una famiglia per l'acquisto di un certo bene e si individui come parametro d'interesse il totale Y di tale spesa nella popolazione di riferimento relativamente ad un certo mese. Tale parametro può essere espresso nella forma

$$Y = \sum_{i=1}^{N} \sum_{j=1}^{G} Y_{ij} \quad ,$$

in cui N è il numero di famiglie della popolazione, G è il numero di giorni nel mese considerato e Y_{ij} è la spesa della famiglia i nel giorno j del mese considerato. Indicando inoltre con

$$Y_{i.} = \sum_{j=1}^{G} Y_{ij}$$

la spesa mensile della famiglia i, si può scrivere il totale Y come

$$Y = \sum_{i=1}^{N} Y_{i.} \quad .$$

Si supponga ora di rilevare la spesa y su un campione di famiglie di numerosità n e di osservare ciascuna famiglia per un numero di giorni pari a g. Si indichi con π_i la probabilità di inclusione della famiglia i nel campione delle famiglie, dipendente dal disegno di campionamento utilizzato, e con τ_j la probabilità di osservazione del giorno j; si osserva che in particolare $\tau_j = g/G$ per ogni j.

A livello di singola famiglia, per stimare $Y_{i.}$, ossia la spesa mensile della famiglia i-esima avendone rilevato le spese per g giorni, si può utilizzare lo stimatore per espansione $\tilde{Y}_{i.}$ definito come

$$\tilde{Y}_{i.} = \sum_{j=1}^{g} \frac{1}{\tau_j} Y_{ij} = \frac{G}{g} \sum_{j=1}^{g} Y_{ij} \quad ,$$

che si può anche scrivere come

$$\tilde{Y}_{i.} = \sum_{j=1}^{G} \frac{1}{\tau_j} d_j Y_{ij} \quad , \qquad (1)$$

essendo d_j una variabile indicatrice che assume valore uno se il giorno j viene rilevato e per la quale vale: $E(d_j) = \tau_j$.

Per stimare invece il totale Y della spesa relativa al mese considerato per tutta la popolazione, si utilizza lo stimatore per espansione

$$\tilde{Y} = \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{j=1}^{g} \frac{1}{\tau_j} Y_{ij} \quad ,$$

che può anche essere espresso nella forma

$$\tilde{Y} = \sum_{i=1}^{N} \frac{1}{\pi_i} \sum_{j=1}^{G} \frac{1}{\tau_j} \delta_{ij} Y_{ij} \quad ,$$

(2)

in cui

$$\delta_{ij} = \begin{cases} 1 & \text{se la famiglia i viene osservata nel giorno j} \\ 0 & \text{altrimenti} \end{cases} \quad .$$

Si osserva che $E(\delta_{ij}) = Prob(\delta_{ij}=1) = \pi_i \tau_j$, dal momento che i due eventi, ovvero la selezione della famiglia i e la selezione del giorno j, sono indipendenti.

Si vuole ora dimostrare che $\tilde{Y}_{i.}$ è uno stimatore corretto per Y_i e che quindi $\tilde{Y}$ è uno stimatore corretto per Y. Infatti, considerando il valore atteso della (1), si ha che

della famiglia i e la selezione del giorno j, sono indipendenti.

Si vuole ora dimostrare che $\tilde{Y}_{i.}$ è uno stimatore corretto per Y_i e che quindi $\tilde{Y}$ è uno stimatore corretto per Y. Infatti, considerando il valore atteso della (1), si ha che

$$E(\tilde{Y}_{i.}) = E\left(\sum_{j=1}^{G} \frac{1}{\tau_j} d_j Y_{ij}\right) = \sum_{j=1}^{G} \frac{1}{\tau_j} Y_{ij} E(d_j) = \sum_{j=1}^{G} \frac{1}{\tau_j} Y_{ij} \tau_j = \sum_{j=1}^{G} Y_{ij} = Y_{i.}$$

,

oppure

$$E(\tilde{Y}_{i.}) = E\left(\sum_{j=1}^{G} \frac{G}{g} d_j Y_{ij}\right) = \sum_{j=1}^{G} \frac{G}{g} Y_{ij} E(d_j) = \sum_{j=1}^{G} \frac{G}{g} Y_{ij} \frac{g}{G} = Y_{i.}$$

;

si può notare che nel caso che venga osservata, per esempio, una settimana (g=7) in un mese di G=30 giorni, il coefficiente di riporto al mese che viene utilizzato, $1/\tau_j$, è pari a 30/7.

Prendendo ora in esame il valore atteso della (2), si ottiene

$$E(\tilde{Y}) = E\left(\sum_{i=1}^{N} \frac{1}{\pi_i} \sum_{j=1}^{G} \frac{1}{\tau_j} \delta_{ij} Y_{ij}\right) = \sum_{i=1}^{N} \frac{1}{\pi_i} \sum_{j=1}^{G} \frac{1}{\tau_j} Y_{ij} E(\delta_{ij})$$

$$= \sum_{i=1}^{N} \frac{1}{\pi_i} \sum_{j=1}^{G} \frac{1}{\tau_j} Y_{ij} \pi_i \tau_j = \sum_{i=1}^{N} \sum_{j=1}^{G} Y_{ij} = Y$$

Pertanto i due stimatori conducono entrambi a stime non distorte.

Ciononostante, è evidente che l'utilizzo di $\tilde{Y}_{i.}$ per riportare al mese le spese rilevate per la famiglia i per un periodo di g giorni non riproduce il comportamento di spesa di tale famiglia se la spesa in esame viene effettuata con frequenza inferiore a 1/g. In altri termini, si consideri ad esempio un bene la cui frequenza di acquisto è una volta ogni due settimane e per il quale viene rilevata la spesa per un periodo di una settimana. A livello di dato elementare, la spesa riportata al mese risulta in ogni caso non veritiera. Infatti, se viene rilevata una spesa non nulla, tale spesa viene moltiplicata per 30/7 attribuendo alla famiglia una spesa doppia di quella che essa in realtà avrebbe effettuato in un mese; se invece viene rilevata una spesa pari a zero, nel riporto al mese tale spesa rimane zero non rispecchiando neanche in questo caso il comportamento del consumatore.

9. Valutazioni dell'effetto della strategia campionaria sulla variabilità delle stime

9.1. Indici di efficienza della strategia di campionamento

Nel presente paragrafo vengono riportate alcune valutazioni degli effetti delle modifiche introdotte a livello di strategia di campionamento, ossia di disegno e di procedura di stima, sulla precisione delle stime prodotte dall'indagine sui consumi. Si osserva a tale proposito che la vecchia indagine, sia per quanto riguarda il disegno di campionamento che per quanto riguarda la procedura di stima, non adottava procedimenti propriamente probabilistici; infatti, il disegno veniva ricavato da quello delle forze di lavoro utilizzando criteri non necessariamente probabilistici e la determinazione del coefficiente di riporto non partiva dalle probabilità d'inclusione derivanti dal disegno, ma scaturiva da un procedimento di tipo matematico che utilizzava totali di popolazione di fonte censuaria. Nel seguito si tenterà comunque di fornire alcune valutazioni di efficienza delle stime prodotte dalla *vecchia* e dalla *nuova* indagine.

E' utile introdurre la seguente simbologia. Indichiamo con d_n e d_v rispettivamente il nuovo e il vecchio disegno di campionamento, con s_n e s_v la nuova e la vecchia procedura di stima, con t_n e t_v la nuova e la vecchia tecnica di raccolta dei dati. Si possono allora indicare le stime del totale Y di una generica voce di spesa, ottenute con la nuova e con la vecchia indagine, rispettivamente come

a. $\tilde{Y}_{d_n t_n s_n}$ *b.* $\tilde{Y}_{d_v t_v s_v}$.

In base ai dati disponibili è stato possibile poi ottenere, applicando rispettivamente ai dati della *nuova indagine* la *vecchia* procedura di stima e ai dati della *vecchia* indagine la *nuova* procedura di stima, le due stime

c. $Y_{d_n t_n s_v}$ *d.* $Y_{d_v t_v s_n}$,

il cui calcolo ha richiesto diversi passaggi, dal momento che è stato necessario, sui dati della

nuova procedura di stima, le due stime

c. $Y_{d_n t_n s_v}$ *d.* $Y_{d_v t_v s_n}$,

il cui calcolo ha richiesto diversi passaggi, dal momento che è stato necessario, sui dati della vecchia indagine, costruire i pesi diretti (che prima non venivano calcolati) per poi applicare il *nuovo* stimatore, mentre sui dati della *nuova* si è dovuto applicare la *vecchia* procedura di stima.

Sulla base delle stime *a-d* si sono potute ottenere delle valutazioni numeriche dell'efficienza delle due strategie in termini di errori di campionamento.

E' necessario premettere però che, dal momento che nella nuova indagine è stata diminuita la numerosità campionaria sia di primo stadio che di secondo stadio, è ragionevole attendersi che gli errori di campionamento della nuova indagine siano più elevati di quelli ottenuti con la vecchia. In conseguenza di ciò, per effettuare un confronto tra la variabilità delle stime prodotte dalle due indagini è necessario eliminare l'influenza della numerosità campionaria. Per ottenere questo risultato si sono messe a confronto la nuova e la vecchia strategia con una *strategia di riferimento* basata sullo stimatore diretto (indicato con s_r) e su un campionamento casuale semplice (indicato con d_r); si sono costruiti così due indici di efficienza, uno per la nuova e uno per la vecchia indagine, ciascuno dei quali è ottenuto come rapporto tra la varianza della stima prodotta dalla strategia in esame (vecchia o nuova) e quella ottenibile dalla strategia di riferimento con un campione di numerosità pari (in termini di famiglie) a quella della strategia in esame. Per il generico parametro d'interesse Y sono stati calcolati i due *effetti della strategia* (nuova e vecchia), ottenuti come

$$EFF(d_n s_n) = \frac{V(Y_{d_n t_n s_n})}{V(Y_{d_r t_n s_r})}$$

$$EFF(d_v t_v) = \frac{\bar{V}(\bar{Y}_{d_v t_v s_v})}{\bar{V}(\bar{Y}_{d_r t_v s_r})},$$

in cui $\bar{V}(\bar{Y}_{d_n t_n s_n})$ è la stima della varianza dello stimatore ottenuto con la nuova strategia e $\bar{V}(\bar{Y}_{d_r t_n s_r})$ è la stima della varianza dello stimatore ottenuto con la strategia di riferimento, basata sullo stimatore diretto e su un campione casuale semplice di numerosità pari a quella della nuova indagine, mentre $\bar{V}(\bar{Y}_{d_v t_v s_v})$ è la stima della varianza dello stimatore ottenuto con la vecchia strategia e $\bar{V}(\bar{Y}_{d_r t_v s_r})$ è la stima della varianza dello stimatore ottenuta con la strategia di riferimento basata sullo stimatore diretto e su un campione casuale semplice con numerosità pari a quella della vecchia indagine.

Come è logico attendersi, le due quantità ora definite assumono valori superiori all'unità, dal momento che un campione a due stadi è in generale meno efficiente di un campione casuale semplice di pari numerosità; quindi, tanto più l'indice di efficienza di una determinata strategia è superiore all'unità, tanto più essa è inefficiente rispetto alla strategia di riferimento. Se ad esempio $EFF(d_n s_n)$ risultasse minore di $EFF(d_v s_v)$, si avrebbe l'indicazione che la nuova strategia è, in un certo senso, più efficiente della vecchia in quanto produce stime la cui precisione si avvicina maggiormente a quella della strategia di riferimento.

9.2. Indici di efficienza del disegno e della procedura di stima

I due indici ora illustrati forniscono una valutazione globale dell'efficienza delle due strategie campionarie; si può pensare di costruire degli indicatori dell'efficienza di aspetti specifici delle strategie, quali il disegno di campionamento e la procedura di stima.

Per quanto riguarda il disegno di campionamento, sono stati calcolati i due *effetti del disegno* (o *deff*) per i due piani campionari in esame (il nuovo e il vecchio); ciascuno di questi due effetti è ottenuto rapportando la varianza della stima diretta ottenuta con il disegno in esame alla varianza della stima diretta ottenuta con il *disegno di riferimento* (campionamento casuale semplice di numerosità pari a quella del disegno in esame). Si ottengono pertanto le espressioni di seguito riportate

$$EFF(d_n) = \frac{\bar{V}(\bar{Y}_{d_n t_n s_r})}{\bar{V}(\bar{Y}_{d_r t_n s_r})}$$

$$EFF(d_v) = \frac{\bar{V}(\bar{Y}_{d_v t_v s_r})}{\bar{V}(\bar{Y}_{d_r t_v s_r})},$$

$$EFF(d_v) = \frac{\bar{V}(\bar{Y}_{d_v t_v s_r})}{\bar{V}(\bar{Y}_{d_r t_v s_r})},$$

in cui $\bar{V}(\bar{Y}_{d_n t_n s_r})$ e $\bar{V}(\bar{Y}_{d_v t_v s_r})$ sono le stime, ottenute rispettivamente con il nuovo e con il vecchio disegno di campionamento, della varianza dello stimatore diretto; mentre $\bar{V}(\bar{Y}_{d_r t_n s_r})$ e $\bar{V}(\bar{Y}_{d_r t_v s_r})$ indicano le stime della varianza dello stimatore diretto ottenute con un campione casuale semplice di numerosità pari rispettivamente a quella della nuova e della vecchia indagine. Questi due indici forniscono una valutazione di quale dei due disegni è meno inefficiente (dal momento che ci si attende che le due quantità siano entrambe superiori all'unità) rispetto al campionamento casuale semplice.

Relativamente alla procedura di stima, è stato calcolato un indice di efficienza del nuovo stimatore rispetto al vecchio. Tale indice si ottiene come rapporto tra la varianza del nuovo e del vecchio stimatore, applicati sullo stesso insieme di dati; di conseguenza, disponendo di due insiemi di dati (della nuova e della vecchia indagine) l'indice in questione può essere calcolato in due modi alternativi utilizzando i dati della nuova o della vecchia indagine, secondo le espressioni di seguito riportate

$$EFF_n(s) = \frac{\bar{V}(\bar{Y}_{d_n t_n s_n})}{\bar{V}(\bar{Y}_{d_n t_n s_v})}, \qquad EFF_v(s) = \frac{\bar{V}(\bar{Y}_{d_v t_v s_n})}{\bar{V}(\bar{Y}_{d_v t_v s_v})}.$$

Nella tabella 1 sono riportati i valori ottenuti per gli indici sopra menzionati per le stime dei totali degli aggregati spesa per generi alimentari, spesa per generi non alimentari e spesa totale, con riferimento alla vecchia e alla nuova indagine per il primo trimestre 1997 e alla nuova per tutto il 1997.

Tabella 1. Indici di efficienza della strategia, del disegno e dello stimatore: confronti tra vecchia e nuova indagine

VECCHIA INDAGINE PRIMO TRIMESTRE 1997					
	Effetto della strategia	Effetto del disegno	Effetto del nuovo stimatore	Effetto del vecchio stimatore	Efficienza nuovo / vecchio stimatore
ALIMENTARI	1.73	1.82	0.93	0.95	0.98
NON ALIMENTARI	1.53	1.63	0.93	0.94	0.99
TOTALE	1.58	1.68	0.93	0.94	0.99
NUOVA INDAGINE PRIMO TRIMESTRE 1997					
	Effetto della strategia	Effetto del disegno	Effetto dello stimatore	Effetto del vecchio stimatore	Efficienza nuovo / vecchio stimatore
ALIMENTARI	1.37	1.45	0.94	1.01	0.93
NON ALIMENTARI	1.30	1.40	0.93	0.94	0.99
TOTALE	1.31	1.42	0.92	0.96	0.96
NUOVA INDAGINE ANNO 1997					
	Effetto della strategia	Effetto del disegno	Effetto dello stimatore	Effetto del vecchio stimatore	Efficienza nuovo / vecchio stimatore
ALIMENTARI	1.39	1.45	0.96	1.01	0.95
NON ALIMENTARI	1.31	1.39	0.94	0.94	1.00
TOTALE	1.32	1.40	0.94	0.96	0.98

Come risulta evidente, il disegno della nuova indagine è sensibilmente più efficiente, o meno inefficiente, del disegno che veniva utilizzato per la vecchia indagine, mentre per quanto riguarda gli stimatori, pur risultando entrambi più efficienti dello stimatore diretto, la differenza di efficienza tra i due è trascurabile.

10. Analisi dell'influenza dei metodi di stima sul livello degli aggregati

10. Analisi dell'influenza dei metodi di stima sul livello degli aggregati

Per fornire una valutazione dell'impatto del metodo di stima sul livello delle stime prodotte dall'indagine, sono state calcolate le quantità *a-d* introdotte nel paragrafo precedente. In particolare, sono state calcolate le stime delle spese medie mensili per generi alimentari, non alimentari e totale utilizzando la *vecchia* e la *nuova* procedura di stima sui dati della *vecchia* e della *nuova* indagine relativi al primo trimestre del 1997. Tali stime sono state calcolate anche per le famiglie classificate secondo: l'ampiezza familiare, la tipologia familiare e la regione.

Nelle tabella 2 è riportato sinteticamente l'impatto che il ridisegno dell'indagine e l'introduzione della nuova procedura di stima hanno determinato sui livelli delle stime delle spese medie per alimentari, non alimentari e totali, calcolate per il primo trimestre 1997 sia sui dati della nuova che sui dati della vecchia indagine.

Tabella 2 - Spese medie ottenute con il vecchio e con il nuovo peso sui dati della nuova e della vecchia indagine per il primo trimestre 1997

	NUOVA INDAGINE		**VECCHIA INDAGINE**		**Differenza totale nuova – vecchia**	**impatto tecnica d'indagine**		**impatto stimatore**	
						n. stim.	**v. stim.**	**n. ind.**	**v. ind.**
	peso vecchio	**peso nuovo**	**peso vecchio**	**peso nuovo**		nuova – vecchia	nuova - vecchia	n. stim. - v. stim.	n. stim. - v. stim.
ALIMENTARI	942.262	914.875	700.120	675.241	214.755	239.634	242.142	- 27.387	- 24.879
NON ALIMENTARI	3.548.472	3.472.168	2.786.576	2.697.978	685.592	774.190	761.896	- 76.304	- 88.598
TOTALE	4.490.734	4.387.042	3.486.696	3.373.219	900.346	1.013.823	1.004.038	- 103.692	- 113.477

E' immediato rilevare che, mentre è molto forte l'effetto della differente rilevazione, è molto più ridotto l'impatto del differente stimatore, anche se comunque le stime ottenute con il *nuovo* stimatore sono inferiori, a parità di dati, a quelle ottenute con il *vecchio* stimatore. Si fa notare che la stima della spesa totale media mensile ottenuta con il peso diretto (cfr. tabelle 2-4) è maggiore (4.577.946) della medesima stima ottenuta con il peso vecchio (4.490.119), che a sua volta è superiore della stima ottenuta con il peso nuovo (4.386.528).

Nelle tabelle 3-5 sono riportate le stime delle medie di spesa rilevate con la nuova indagine classificate per regione geografica, ampiezza e tipologia familiare, ottenute con il vecchio e con il nuovo stimatore e, per confronto, le medesime stime calcolate utilizzando il peso diretto.

Tabella 3 – Spese medie per regione ottenute con il vecchio peso, con il nuovo e con il peso diretto sui dati della nuova indagine per il primo trimestre 1997

	Spese alimentari			**spese non alimentari**			**spesa totale**		
Regione	**peso vecchio**	**peso nuovo**	**peso diretto**	**peso vecchio**	**peso nuovo**	**peso diretto**	**peso vecchio**	**peso nuovo**	**peso diretto**
PIEMONTE	914.279	882.977	930.667	3.463.192	3.359.293	3.515.927	4.377.471	4.242.270	4.446.594
VALLE D'A.	874.124	857.563	899.643	3.772.768	3.691.895	3.859.268	4.646.892	4.549.457	4.758.911
LOMBARDIA	1.025.170	995.278	1.039.371	4.270.905	4.187.920	4.362.623	5.296.075	5.183.199	5.401.993
BOLZANO	797.848	804.148	821.101	4.253.290	4.275.503	4.375.173	5.051.138	5.079.652	5.196.273
TRENTO	911.972	865.899	922.946	3.974.932	3.867.231	4.120.893	4.886.904	4.733.129	5.043.840
VENETO	893.889	876.754	908.693	4.281.225	4.288.905	4.382.751	5.175.114	5.165.659	5.291.444
FRIULI V. G.	826.773	804.085	869.768	3.199.426	3.020.045	3.270.391	4.026.199	3.824.129	4.140.158
LIGURIA	909.782	870.478	885.638	3.115.867	3.054.901	3.073.301	4.025.650	3.925.380	3.958.939
EMILIA R.	914.220	892.613	906.109	4.226.672	4.117.005	4.145.535	5.140.893	5.009.618	5.051.643
TOSCANA	997.321	943.052	988.545	3.680.156	3.529.954	3.657.315	4.677.476	4.473.006	4.645.860
UMBRIA	956.655	907.236	945.348	3.785.666	3.543.973	3.730.680	4.742.322	4.451.209	4.676.028
MARCHE	996.468	982.125	1.019.125	3.475.731	3.371.880	3.535.306	4.472.199	4.354.005	4.554.431
LAZIO	978.304	953.223	981.348	3.585.559	3.512.690	3.572.301	4.563.863	4.465.913	4.553.649
ABRUZZI	891.818	865.845	964.093	3.604.232	3.459.644	3.964.890	4.496.050	4.325.489	4.928.983
MOLISE	823.486	849.968	876.626	2.758.771	2.792.905	3.002.344	3.582.257	3.642.872	3.878.970
CAMPANIA	986.337	959.622	991.091	3.117.791	3.087.007	3.224.713	4.104.128	4.046.629	4.215.804
PUGLIA	923.134	897.885	971.363	2.939.581	2.866.239	3.166.207	3.862.715	3.764.124	4.137.570
BASILICATA	897.494	894.026	958.666	2.479.608	2.449.265	2.699.792	3.377.102	3.343.291	3.658.458
CALABRIA	988.928	941.127	1.024.321	2.890.521	2.833.299	2.977.622	3.879.450	3.774.425	4.001.943
SICILIA	833.825	815.788	869.894	2.351.029	2.258.255	2.437.850	3.184.854	3.074.043	3.307.743

PUGLIA	923.134	897.885	971.363	2.939.581	2.866.239	3.166.207	3.862.715	3.764.124	4.137.570
BASILICATA	897.494	894.026	958.666	2.479.608	2.449.265	2.699.792	3.377.102	3.343.291	3.658.458
CALABRIA	988.928	941.127	1.024.321	2.890.521	2.833.299	2.977.622	3.879.450	3.774.425	4.001.943
SICILIA	833.825	815.788	869.894	2.351.029	2.258.255	2.437.850	3.184.854	3.074.043	3.307.743
SARDEGNA	883.112	862.917	893.929	3.355.558	3.416.335	3.535.108	4.238.670	4.279.252	4.429.037
ITALIA	**942.192**	**914.839**	**957.361**	**3.547.927**	**3.471.689**	**3.620.135**	**4.490.119**	**4.386.528**	**4.577.496**

Tabella 4 – Spese medie per ampiezza familiare ottenute con il vecchio peso, con il nuovo e con il peso diretto sui dati della nuova indagine per il primo trimestre 1997

	Spese alimentari			**spese non alimentare**			**spesa totale**		
Ampiezza familiare	**peso vecchio**	**peso nuovo**	**peso diretto**	**peso vecchio**	**peso nuovo**	**peso diretto**	**peso vecchio**	**peso nuovo**	**peso diretto**
1	576.346	583.031	585.296	2.270.463	2.291.515	2.322.006	2.846.809	2.874.545	2.907.301
2	828.952	830.678	835.079	3.120.647	3.141.084	3.179.007	3.949.599	3.971.762	4.014.086
3	1.030.031	1.022.423	1.032.159	4.034.423	4.021.999	4.038.414	5.064.454	5.044.423	5.070.573
4	1.137.218	1.124.200	1.133.519	4.281.575	4.203.784	4.235.428	5.418.793	5.327.984	5.368.946
5	1.296.201	1.289.649	1.288.022	4.483.345	4.460.219	4.465.271	5.779.546	5.749.869	5.753.293
6 e più	1.369.655	1.382.183	1.356.220	4.375.138	4.522.350	4.422.546	5.744.793	5.904.534	5.778.766
Totale	**942.192**	**914.839**	**957.361**	**3.547.927**	**3.471.689**	**3.620.135**	**4.490.119**	**4.386.528**	**4.577.496**

Tabella 5 – Spese medie per tipologia familiare ottenute con il vecchio peso, con il nuovo e con il peso diretto sui dati della nuova indagine per il primo trimestre 1997

	spese alimentari			**spese non alimentari**			**spesa totale**		
Tipologia familiare	**peso vecchio**	**peso nuovo**	**peso diretto**	**peso vecchio**	**peso nuovo**	**peso diretto**	**peso vecchio**	**peso nuovo**	**peso diretto**
PERSONA SOLA<35	600.589	617.457	612.124	3.928.249	4.032.234	4.093.614	4.528.838	4.649.691	4.705.738
PERSONA SOLA 35-64	632.945	644.998	648.951	3.100.770	3.149.583	3.171.755	3.733.715	3.794.581	3.820.706
PERSONA SOLA>65	540.525	543.620	545.286	1.508.666	1.515.623	1.529.734	2.049.191	2.059.243	2.075.020
COPPIA SENZA FIGLI P.R.<35	729.690	730.000	736.543	3.655.962	3.692.925	3.767.081	4.385.652	4.422.925	4.503.624
COPPIA SENZA FIGLI P.R. 35-64	846.729	855.265	856.511	3.836.832	3.907.093	3.967.933	4.683.561	4.762.358	4.824.445
COPPIA SENZA FIGLI P.R. >65	865.065	866.631	870.017	2.531.130	2.547.268	2.573.512	3.396.195	3.413.899	3.443.529
COPPIA CON UN FIGLIO	1.023.907	1.014.846	1.024.287	4.126.040	4.119.446	4.122.922	5.149.947	5.134.292	5.147.209
COPPIA CON DUE FIGLI	1.134.294	1.124.908	1.130.467	4.274.480	4.220.553	4.223.331	5.408.774	5.345.461	5.353.798
COPPIA CON TRE E PIU FIGLI	1.303.918	1.303.351	1.295.777	4.406.634	4.439.968	4.401.284	5.710.552	5.743.319	5.697.061
MONOGENITORE CON ALMENO UN FIGLIO MIN.	958.620	949.511	979.912	3.044.142	3.012.119	3.120.025	4.002.762	3.961.630	4.099.937
MONOGENITORE CON FIGLI MAGGIOR.	910.298	915.839	919.816	3.265.666	3.270.823	3.258.395	4.175.963	4.186.662	4.178.211
ALTRE TIPOLOGIE	1.102.337	1.041.772	1.095.792	4.021.588	3.816.066	4.009.972	5.123.925	4.857.838	5.105.764
Totale	**942.192**	**914.839**	**957.361**	**3.547.927**	**3.471.689**	**3.620.135**	**4.490.119**	**4.386.528**	**4.577.496**

Dall'esame della tabella 3 viene confermato a livello regionale quanto emerso a livello nazionale; infatti, le stime ottenute con il nuovo peso sono quasi uniformemente inferiori a quelle ottenute con il peso vecchio e il peso diretto (salvo in alcune regioni minori, come Bolzano e il Molise).

Per quanto riguarda invece le medie di spesa calcolate per ampiezza e tipologia familiare, si nota come l'uniformità di comportamento venga meno: infatti sulle famiglie più piccole, uno o due componenti, le stime ottenute con il nuovo peso sono leggermente più elevate di quelle ottenute con il peso diretto e con il peso vecchio. Sebbene tali differenze siano di entità contenuta, è comunque difficile valutare l'impatto della differente post-stratificazione in quanto la vecchia procedura di stima non partiva dalla costruzione dei pesi diretti.