ANNALI DI STATISTICA

Anno 120 Serie IX - Vol. 10

ATTI
DELLA GIORNATA DI STUDIO
SUL CAMPIONAMENTO STATISTICO

ROMA, 27 Aprile 1989

ISTITUTO NAZIONALE DI STATISTICA
ROMA 1991

L'Istat autorizza la riproduzione parziale o totale del contenuto del presente volume con la citazione della fonte.

Supplemento all'Annuario Statistico Italiano

ISSN: 0390-6434

EDIGRAFITAL S.p.A. - Teramo - Lettera ordinativo n. 9254 del 29-4-89 - Copie 1.200

INDICE

INTERVENTI

PRESENTAZIONE

La più recente attività dell'Istituto si è fra l'altro caratterizzata per lo sviluppo delle indagini campionarie con il quale si è dato seguito ad un preciso impegno programmatico più volte ribadito nel corso degli ultimi anni.

Questa maggiore attenzione, nei confronti di un'area così nevralgica per la produzione dell'Istat, ha in particolare comportato una più puntuale valutazione delle performance dei piani di campionamento nel tempo predisposti per la realizzazione delle varie rilevazioni parziali e la messa a punto di nuovi disegni sollecitata dagli obiettivi via via assunti dai programmi annuali.

A supporto di questo rinnovato impegno si è naturalmente reso necessario intensificare l'attività di ricerca in campo metodologico che ha avuto due principali sbocchi: la ristrutturazione di piani di campionamento da tempo operanti, da un lato, e, dall'altro, l'individuazione di schemi di riferimento alternativi soprattutto per le esigenze poste dall'esecuzione di indagini che si volevano snelle, più rapide e perciò basate su universi di più ridotte dimensioni.

L'attività di ricerca portata avanti dal Reparto Studi dell'Istituto ha potuto al riguardo avvalersi del contributo fornito da un'apposita Commissione di studio nella quale hanno svolto un ruolo attivo eminenti studiosi di estrazione accademica. Essa è stata la sede di un interessante confronto di idee fra i ricercatori dell'Istituto e quelli impegnati in ambito universitario.

Con il Convegno si è voluto in qualche modo offrire una testimonianza del lavoro svolto ed allargare la discussione ad una più vasta platea di esperti d'area.

Con gli «Atti» che ora vengono stampati ci si augura che il confronto possa avere un ulteriore seguito e che si possa consolidare, in vista dei nuovi obiettivi, quello spirito di sempre più stretta collaborazione che ha così positivamente influito sui lavori della «Giornata di studio».

RELAZIONI

RELAZIONE SULL'ATTIVITÀ DELLA COMMISSIONE

di *Giuseppe Leti*

Nel 1983 il Presidente dell'Istat, prof. G.M. Rey — che ringrazio, insieme al prof. V. Siesto, per avermi invitato ad introdurre i lavori di questa Giornata di studio — costituì una Commissione col compito di formulare proposte in merito alla progettazione ed applicazione di campioni per le indagini campionarie dell'Istituto. La Commissione, presieduta da me, quale membro del Consiglio Superiore di Statistica, era costituita dai funzionari dell'Istat: R. Angeloni, M. Masselli, A. Russo e F. Zannella, e dai membri esterni dell'Istat: L. Bergonzini, L. Fabbris, A. Herzel. Dopo circa un biennio L. Bergonzini si ritirò ed entrò a far parte della Commissione G. Cicchitelli; ai funzionari dell'Istat si sono aggiunti nel tempo M. Di Traglia e F. Crescenzi.

Compito della Commissione, individuato negli incontri con i Dirigenti dell'Istituto, era quello di formulare, in cooperazione con il Reparto Studi, proposte operative su:

1. l'organizzazione di un sistema di campioni dell'Istat che tenga conto di diverse esigenze di rappresentatività territoriale e di diverse necessità di tempestività;

2. la determinazione di disegni campionari specifici e/o di un master sample;

3. l'organizzazione e la standardizzazione delle indagini campionarie in tutte le loro fasi, dalla loro progettazione alla raccolta dei dati ed alla elaborazione dei dati raccolti;

4. il controllo della qualità dei risultati campionari;

5. la promozione di una «cultura» campionaria all'interno dell'Istituto.

La Commissione, per assolvere l'incarico affidatole, ha anzitutto fatto una ricognizione sullo stato e sull'organizzazione delle indagini campionarie effettuate dall'Istat, con un'attenzione particolare alle metodologie e ai piani di campionamento utilizzati, all'analisi effettuata sulla qualità dei dati e alla determinazione degli errori non campionari. La Commissione ha anche ritenuto opportuno do-

cumentarsi sulle indagini campionarie, condotte da alcuni Istituti esteri di statistica, studiando in particolare la loro organizzazione.

Si rilevò che nel 1982 solo il 10% delle indagini effettuate era costituito da indagini campionarie, ossia in totale 17 indagini campionarie, di cui 4 rivolte a privati cittadini, 8 ad imprese e 5 ad amministrazioni pubbliche. La stessa situazione sussisteva nel 1985.

Oggi, passati altri 4 anni si ha la seguente situazione:

29 indagini campionarie su un totale di 200 indagini effettuate dall'Istat e di esse 7 riguardano i privati cittadini e 22 le imprese o simili. È nella fase iniziale l'indagine multiscopo sulle famiglie.

In concomitanza, si sono raggiunti buoni standard per quanto concerne la predisposizione e l'analisi dei piani di campionamento. In particolare:

— si è invertita la tendenza a ricorrere al campione base della indagine sulle forze di lavoro per le rilevazioni di altra natura: attualmente per le differenti indagini vengono predisposti sempre più frequentemente piani campionari ad hoc;

— si provvede, in modo sistematico, al calcolo degli errori campionari delle stime e degli indicatori sull'efficienza della stratificazione;

— vengono utilizzate tecniche di post-stratificazione;

— si sta affrontando il problema dell'analisi delle sostituzioni e della presentazione degli errori di campionamento.

Però permangono per le indagini condotte dall'Istituto alcuni elementi di perplessità, quali:

— non si tiene conto in modo congiunto degli elementi costitutivi della spesa e di quelli degli errori (campionari e non campionari);

— la famiglia viene spesso scelta come unità di rilevazione anche quando sarebbe preferibile rilevare direttamente gli individui;

— le basi campionarie, soprattuto per le indagini sulle famiglie, presentano problemi per quanto riguarda la completezza ed il rinnovo;

— la mancanza di riferimento ad unità territoriali diverse dai comuni (USL, concentrazioni urbane, aree, etc.);

— l'insufficiente sperimentazione su tecniche di rilevazione alternative (per aree, retrospettive, longitudinali, etc.).

La Commissione ha proceduto dapprima ad una classificazione della tipologia delle indagini campionarie ed ha quindi individua-

to le fasi in cui si articola la predisposizione della rilevazione e del controllo di qualità; su tali basi ha formulato un piano coerente di sperimentazione e predisposizione di metodologie che ha proposto all'Istituto. Il piano è stato predisposto in collaborazione con i progetti «qualità dei dati» e «campioni» del Reparto Studi ed ha implicato ed implica il loro coinvolgimento per la sua attuazione.

In particolare si è lavorato: alla predisposizione di un archivio informatizzato di dati censuri (e dei programmi di accesso) da utilizzare per la stratificazione dei comuni italiani, all'analisi dei problemi e delle metodologie per la stratificazione (con l'approntamento di procedure informatiche sperimentali), allo studio dei criteri di selezione delle unità di primo stadio e delle proprietà degli stimatori mediante procedimenti di simulazione, ed infine è stato affrontato, anch'esso mediante procedure di simulazione, il campionamento areale.

Sulla base delle esperienze condotte nell'ambito del Reparto Studi e delle risorse disponibili, la Commissione ha ritenuto che l'attività di ricerca e studio dovesse concentrarsi, prioritariamente, sulle indagini sulla popolazione:

a) mettendo a punto una metodologia generalizzata per le indagini campionarie;

b) predisponendo la relativa procedura informatica;

c) predisponendo le normative e compilando manuali teorico-pratici ad uso dei reparti interessati;

d) aggiornando il personale.

La decisione di prendere in esame gli aspetti metodologici relativi alle indagini campionarie sulle popolazioni è stata motivata da una serie di considerazioni che possono così essere sintetizzate:

I) costituiscono una quota importante delle indagini campionarie attualmente effettuate dall'Istituto;

II) presentano una notevole omogeneità sia delle modalità di rilevazione che degli schemi di campionamento adottati;

III) sono concentrate nella quasi totalità in un unico Reparto dell'Istituto;

IV) costituiscono il gruppo di indagini omogenee sulle quali, in maggior misura, sono state già avviate, all'interno del Reparto Studi, sia una riflessione metodologica, sia studi sperimentali sui diversi aspetti del disegno campionario.

Prima di sintetizzare i risultati raggiunti nell'ambito della Commissione è da precisare che gli studi effettuati hanno riguardato indagini campionarie di media e grande taglia sulla popolazione; e che sono stati condotti sulla base di sperimentazioni su tre regioni, (Piemonte, Toscana e Calabria che possono ritenersi «tipiche» per le ricerche in questione) e su alcune variabili considerate rilevanti per le indagini Istat.

I problemi inerenti la stratificazione sono stati affrontati da Francesco Zannella, che ha sperimentato 11 differenti criteri di formazione degli strati. Tali criteri sono parte integrante di una procedura generalizzata di stratificazione dei comuni (predisposta dallo stesso Zannella) che permette all'utilizzatore di ottenere rapidamente una suddivisione dei comuni in strati, consentendo una notevole flessibilità nelle varie fasi del processo e nello stesso tempo fornendo come output una serie di informazioni utili per la scelta del procedimento migliore sia dal punto di vista dell'efficienza che della praticabilità. Infatti la scelta di un metodo di stratificazione non può essere basata esclusivamente su considerazioni relative alla efficienza, ma deve tener conto anche della sua praticabilità, e questa è spesso legata al numero dei comuni e alla popolazione residente negli strati formati. La stratificazione basata sulla popolazione residente presenta, per quanto riguarda gli aspetti legati alla praticabilità, notevoli vantaggi, dei quali uno è la possibilità di essere aggiornata annualmente, a differenza delle altre variabili la cui disponibilità, a livello comunale, è legata ai censimenti, e che quindi possono risultare obsolete nell'intervallo intercensuario. Inoltre la stratificazione basata sulla popolazione residente non risulta meno efficiente di quelle basate su altre variabili, a meno che queste ultime non coincidano con le variabili oggetto di indagine o non siano con queste fortemente correlate.

Abitualmente l'Istat utilizza, per le indagini sulla popolazione, il campionamento a più stadi stratificato, con selezione al primo stadio di un comune per strato con probabilità proporzionali alla dimensione. A favore della selezione di un solo comune, si possono portare vari argomenti di natura pratica e teorica; essa infatti, non solamente tende a ridurre la varianza, ma nell'ambito dei criteri di selezione casuale, è il metodo che introduce il controllo più profondo nel processo di individuazione dei comuni campione. Tuttavia, alcune possibilità, legate alla selezione di due o più comuni per strato (ad es. la stima corretta della varianza e la compenetrazione del cam-

pione per valutare l'errore di risposta), hanno portato ad esaminare la convenienza di tali disegni campionari. Dopo varie prove effettuate da Luigi Fabbris, si è ritenuto opportuno concentrare l'attenzione su 11 tecniche di selezione delle unità di primo stadio. L'analisi dei risultati della sperimentazione portano a ritenere consigliabili, al fine di formare campioni «generalizzati» per indagini svolte in una sola occasione, l'utilizzo della tecnica sistematica (nel testo tecnica numero 2) o quella senza reimmissione (tecnica numero 4) per campioni di qualsiasi numerosità e la tecnica senza reimmissione con lo stimatore di Murthy (tecnica numero 8) per campioni non superiori a 3.

In base alle analisi ed alle sperimentazioni effettuate, è stato delineato da Fabbris e Zannella uno schema di procedura generalizzata per la stratificazione e la selezione dei comuni nelle indagini campionarie sulla popolazione. Tale procedura consentirà all'utilizzatore di estrarre, rapidamente e con semplici istruzioni, un campione di comuni mediante un disegno campionario a due stadi con stratificazione delle unità di primo stadio e con probabilità di selezione variabili. Essa, inoltre, consentirà il calcolo delle varianze campionarie delle stime di frequenze relative o assolute di caratteristiche della popolazione, rilevate al censimento e considerate come oggetto di indagine.

Le relative procedure, utilizzate in via sperimentale, vanno ancora riviste ed implementate da un esperto informatico secondo le indicazioni del documento di Fabbris e Zannella, contenuto nel presente volume.

Nella procedura generalizzata la numerosità di primo e secondo stadio, costituisce una informazione esterna mentre anch'essa dovrebbe essere determinata mediante una procedura specifica.

La Commissione, su indicazione del Presidente dell'Istat, ha iniziato lo studio della determinazione della numerosità campionaria di primo e secondo stadio e già sono stati prodotti approfonditi documenti in merito, ad opera di Aldo Russo, Stefano Falorsi, Fabio Crescenzi e Vittoria Buratta.

Oltre al completamento degli studi volti alla determinazione della numerosità campionaria di primo e secondo stadio, rimangono da affrontare i seguenti problemi:

1) implementazione informatica della procedura generalizzata individuata;

2) predisposizione di un manuale d'uso della stessa;

3) analisi e selezione dei procedimenti di stima anche in relazione agli schemi campionari trattati.

Altri sviluppi che sono emersi dalle discussioni nell'ambito della Commissione riguardano:

a) la determinazione della procedura per il calcolo dell'errore di campionamento;

b) lo studio degli errori non campionari;

c) lo studio del problema dei «piccoli campioni».

Prima di passare ad esporre un altro punto che è stato oggetto di studio da parte della Commissione è da segnalare che i tre documenti predisposti nell'ambito della Commissione e sottoposti all'attenzione dei partecipanti a questa giornata di studio, possono anche riflettere opinioni e valutazioni dell'autore.

Un altro problema che si è posto la Commissione è quello del ricorso all'uso dei campioni areali nel campo delle rilevazioni campionarie sulla popolazione e sulle imprese (o sulle unità locali) industriali, commerciali e del settore terziario, nel quale l'Istat non ha mai utilizzato il campionamento areale. La necessità di prendere in considerazione un tale tipo di campionamento è nata dalle difficoltà relative ad un rapido ed efficace aggiornamento degli schedari delle imprese. Il campionamento areale è infatti preferibile a quello basato sullo schedario delle unità del campo di osservazione quando lo schedario non è completo e ben aggiornato e subisce rapide variazioni. Vantaggi del campione areale sono, tra l'altro, i seguenti:

a) le aree (opportunamente piccole) potrebbero essere stratificate ed il campione sarebbe ben diffuso;

b) non si dovrebbe procedere a sostituzioni.

Un primo studio è stato quello di considerare come aree le sezioni di censimento che, a parere unanime della Commissione, almeno per quanto concerne rilevazioni sulle unità locali industriali, commerciali e del settore terziario (e, ovviamente, sulle imprese), non si prestano come basi di campione areolari, se non altro per l'eccessiva variabilità del numero di unità locali contenute in ciascuna sezione. Ciò del resto era da aspettarsi dato che le sezioni di censimento sono state formate per rispondere ad esigenze diverse da quelle cui dovrebbe rispondere una buona base areolare.

Nell'ambito della Commissione è stata quindi proposta una indagine sperimentale estesa ad un gruppo di comuni di varie dimensioni allo scopo di verificare concretamente quali difficoltà possono insorgere sul piano pratico con riguardo alla formazione di aree secondo criteri specifici ed alla rilevazione stessa, sia in termini organizzativi che esecutivi e di costo.

L'attività della Commissione ha ovviamente coperto solo in piccola parte l'opera dell'Istat nel campo delle indagini campionarie, opera che riguarda sia la tecnica campionaria sia l'analisi della qualità dei risultati del campionamento. Aldo Russo e Mauro Masselli esporranno quanto è stato fatto in merito nei progetti da essi diretti. Aldo Russo esporrà i risultati già ottenuti dal Progetto di Studio dei campioni e si è occupato tra l'altro dei seguenti problemi:

1. tecniche speciali di stima per piccoli domini territoriali;

2. calcolo della varianza campionaria e presentazione mediante modelli degli errori standard;

3. stimatori utilizzati per la determinazione delle stime oggetto d'indagine;

4. utilizzazioni di informazioni ausiliarie ai fini della formazione del campione;

5. analisi del Design Effect.

Mauro Masselli, che dirige il Progetto di Studio sulla qualità dei dati, esporrà quanto nel progetto si è raggiunto con particolare riguardo:

1. alla qualità dell'informazione statistica;

2. al sistema di controllo dell'indagine;

3. alla prevenzione dell'errore;

4. alla correzione dell'errore;

5. alla stima dell'errore.

L'opera della Commissione ha avuto anche — come ha sottolineato il prof. Siesto — un effetto indiretto sull'Istituto avendo migliorato la cultura campionaria nell'ambito dell'Istituto stesso. È questo un risultato assai importante che forse è in parte dovuto alla diaspora di personale dell'Ufficio Studi nei vari reparti operativi, diaspora che, se da un lato ha arricchito di esperienze tali reparti, dall'altro

ha talvolta condizionato il lavoro della Commissione che non sempre ha potuto contare sul sussidio dello stesso nucleo di ricercatori.

Prima di inoltrarsi nei lavori della Giornata, desidero rivolgere un vivo ringraziamento al Presidente dell'Istituto e a quanti nell'Istat hanno portato la loro opera in supporto di quella della Commissione.

METODOLOGIA, PROGRAMMI E SPERIMENTAZIONI RELATIVI ALLA PROGETTAZIONE DI UNA PROCEDURA GENERALIZZATA PER LA STRATIFICAZIONE DEI COMUNI

di *Francesco Zannella*

1. PREMESSA

Come è noto i piani di campionamento utilizzati nell'Istat per effettuare le rilevazioni campionarie sulla popolazione sono, generalmente, a due stadi con stratificazione delle unità di primo stadio (i comuni) e scelta sistematica, o mediante estrazione casuale semplice, delle unità di secondo stadio (le famiglie) che, a seconda del tipo d'indagine, possono rappresentare le unità elementari di rilevazione o grappoli di unità elementari (gli individui).

Il ricorso ad un campionamento a due stadi è motivato, più che da considerazioni di natura metodologica, da esigenze di carattere organizzativo ed economico derivanti:

a) dalla suddivisione amministrativa del territorio nazionale in comuni e dall'esistenza presso questi ultimi delle anagrafi della popolazione, che costituiscono le liste dalle quali vengono sorteggiate le unità finali di rilevazione;

b) dalla necessità di ridurre i costi di rilevazione contenendo la dispersione territoriale delle famiglie da intervistare mediante la loro concentrazione su un numero limitato di comuni.

La stratificazione dei comuni viene introdotta per diverse ragioni teoriche e pratiche:

a) per migliorare l'efficienza del disegno campionario, in quanto i comuni presentano generalmente una elevata variabilità dei caratteri oggetto di rilevazione e una loro suddivisione in gruppi più omogenei consente di ottenere stimatori più precisi;

b) per poter programmare il campione a livelli territoriali disaggregati; infatti spesso tra gli obiettivi di un'indagine c'è quello di fornire stime di una prefissata precisione per particolari sottogruppi della

popolazione (domini di studio), come le regioni, le ripartizioni geografiche o le classi di ampiezza demografica dei comuni;

c) per tener conto delle esigenze amministrative ed organizzative che richiedono il raggruppamento territoriale dei comuni in regioni o province, in modo da demandare agli uffici di statistica competenti per territorio la supervisione delle operazioni connesse con l'implementazione del campione e l'esecuzione della rilevazione stessa (estrazione e formazione degli elenchi delle famiglie da intervistare, istruzioni e controllo dei rilevatori, raccolta e prima revisione dei questionari, ecc.).

Le fasi attraverso le quali si perviene alla formazione del campione da utilizzare per una determinata indagine possono essere così sintetizzate:

- definizione dei domini di studio;
- individuazione delle variabili più rilevanti da considerare come «variabili guida» sia per la determinazione della numerosità del campione che per la formazione degli strati;
- scelta dei livelli attesi di precisione delle stime in corrispondenza di ciascun dominio di studio;
- scelta del numero di comuni da estrarre da ciascuno strato e della tecnica di selezione dei comuni;
- calcolo delle numerosità campionarie in primo e in secondo stadio per ogni dominio di studio;
- individuazione dei comuni autorappresentativi, ossia dei comuni che formano strato a se stante;
- stratificazione dei comuni non autorappresentativi;
- ripartizione tra gli strati del campione di secondo stadio;
- selezione dei comuni campione in base alla tecnica prescelta;
- estrazione da ciascun comune campione delle unità di secondo stadio.

Il percorso logico che viene seguito per la progettazione del campione è in genere sempre lo stesso, mentre le scelte che vengono effettuate e le soluzioni date ai problemi che via via si presentano dipendono dagli obiettivi e dai vincoli amministrativi ed organizzativi dell'indagine per la quale il campione deve essere programmato.

Partendo da queste considerazioni la «Commissione per la progettazione e l'applicazione dei campioni» istituita presso l'Istat, si è posta l'obiettivo di approntare un disegno campionario a due stadi con stratificazione delle unità di primo stadio, generalizzato e completamente informatizzato, da utilizzare per le indagini che l'Istat deve condurre sulla popolazione.

Nella predisposizione del disegno generalizzato potevano essere seguite due diverse strategie:

a) progettazione e selezione di un campione guida (master sample) di grandi dimensioni, dal quale estrarre di volta in volta campioni di numerosità variabile a seconda delle esigenze conoscitive dell'indagine;

b) progettazione di una procedura flessibile rispetto alle scelte che devono essere effettuate nelle diverse fasi della programmazione del campione, in modo da poter adottare soluzioni ad hoc per ogni specifica indagine.

Tenuto conto che attualmente l'Istat effettua numerose indagini campionarie sulla popolazione con finalità conoscitive estremamente diversificate (occupazione, consumi, salute, letture, strutture e comportamenti familiari, ecc.) e che il loro numero è certamente destinato ad aumentare nei prossimi anni, la Commissione ha ritenuto più conveniente percorrere quest'ultima strada, in quanto la soluzione basata sul «master sample» avrebbe portato a campioni poco efficienti rispetto agli obiettivi specifici delle singole indagini.

L'attività della Commissione si è concentrata in una prima fase sui problemi relativi alla stratificazione e alla selezione dei comuni, rimandando ad un secondo tempo quelli inerenti la determinazione della numerosità campionaria.

In questo lavoro, dopo aver esaminato i criteri attualmente utilizzati dall'Istat per la stratificazione dei comuni, viene descritta e sperimentata una procedura di stratificazione che, una volta implementata, dovrebbe consentire all'utilizzatore di ottenere, mediante semplici istruzioni e in tempi relativamente brevi, una stratificazione dei comuni con la possibilità di poter scegliere in ogni fase tra diverse soluzioni.

2. I PROCEDIMENTI DI STRATIFICAZIONE ATTUALMENTE UTILIZZATI

La metodologia per la stratificazione dei comuni, e più in generale quella per la programmazione di un piano di campionamento, è stata per lungo tempo ancorata al primo disegno campionario per indagini sulla popolazione messo a punto attorno alla metà degli anni '50 per le rilevazioni delle forze di lavoro (Istat 1958, 1969, 1978). Il piano di campionamento, predisposto per fornire stime a livello provinciale, utilizza come basi territoriali per la stratificazione i settori statistici, ottenuti raggruppando comuni contigui all'interno di ciascuna provincia. I settori statistici sono stati formati, sulla base dei risultati del censimento del 1951, in modo da costituire aree omogenee dal punto di vista economico e non troppo variabili per numero di comuni, popolazione residente e superficie (Istat 1958a).

All'interno di ciascun settore statistico i comuni sono stati suddivisi sulla base della popolazione residente in:

a) comuni con almeno 20.000 abitanti (autorappresentativi), ciascuno dei quali forma uno strato a sé stante;

b) i comuni con meno di 20.000 abitanti (non autorappresentativi) che vengono ulteriormente stratificati per zona altimetrica ed attività economica prevalente, valutata sulla base delle percentuali di popolazione attiva impegnata nei tre settori (agricoltura, industria, altre attività).

I comuni sui quali viene condotta l'indagine sono costituiti da tutti i comuni autorappresentativi e da un campione di quelli non autorappresentativi, ottenuto selezionando un comune per strato con probabilità proporzionale alla popolazione residente.

La numerosità del campione in primo stadio dipende quindi dalla soglia che viene fissata per la determinazione dei comuni autorappresentativi e dal numero degli strati in cui vengono raggruppati i restanti comuni, numero che non viene stabilito a priori ma che risulta determinato dallo stesso criterio di stratificazione adottato. Infatti, la numerosità degli strati è data dal numero delle diverse attività economiche prevalenti che si riscontrano nelle zone altimetriche presenti in ciascun settore statistico.

Così se in un settore statistico si hanno soltanto comuni di montagna e di collina (due zone altimetriche) e nei comuni di montagna

l'attività prevalente è sempre l'agricoltura (una sola attività economica prevalente) e in quelli di collina prevale in alcuni comuni l'industria e in altri il terziario (due attività economiche prevalenti) si avranno in totale tre strati.

È evidente che un tale modo di procedere non consente di tenere sotto controllo il numero degli strati che si formano, e quindi il numero dei comuni che devono essere campionati. Pertanto il metodo risulta poco flessibile rispetto alle variazioni della numerosità nel primo stadio di campionamento. Nel caso sia necessario aumentare il numero dei comuni campione, come spesso avviene per soddisfare le esigenze conoscitive di regioni e province, occorre introdurre un qualche criterio (ad esempio la popolazione residente) per un'ulteriore suddivisione degli strati formati in precedenza.

Inoltre il numero di comuni e l'ammontare della popolazione possono risultare molto variabili da strato a strato, così come una forte variabilità si può riscontrare tra le popolazioni dei comuni che appartengono allo stesso strato. Ciò può comportare da un lato una riduzione dell'efficienza del disegno campionario conseguente alla notevole variabilità delle probabilità di selezione dei comuni, e dall'altro problemi di praticabilità, in quanto può accadere che in uno strato siano presenti comuni con una popolazione residente inferiore a quella che deve essere campionata.

Un'ultima considerazione che deve essere svolta riguarda l'efficienza del procedimento di stratificazione. Diverse sperimentazioni effettuate nell'ambito della Commissione Istat per gli studi statistici ed econometrici interessanti la programmazione economica (Biggeri ed altri 1977, Zani e Sicuri 1977a, 1977b) hanno evidenziato come il raggruppamento dei comuni all'interno del settore statistico e della zona altimetrica in base al criterio dell'attività economica prevalente non necessariamente costituisce il metodo migliore per ottenere strati più omogenei rispetto alle percentuali di popolazione attiva per settore di attività economica.

Occorre aggiungere che il procedimento di stratificazione adottato prevede un'ulteriore suddivisione degli strati che presentano un campo di variazione della percentuale di addetti all'attività prevalente superiore a 25%. Ma ciò, oltre a rendere più macchinoso il metodo, non garantisce il contenimento delle variabilità per le percentuali di addetti alle altre due attività economiche.

Questo piano di campionamento ha costituito per molti anni lo schema di riferimento per la progettazione dei campioni impiegati anche per altre indagini sulla popolazione, che hanno riguardato i fenomeni più diversi come i consumi, le letture, le vacanze, la salute, ecc.

L'utilizzazione dello stesso disegno campionario, e a volte dello stesso campione, era giustificata dalle difficoltà, più operative che metodologiche, che s'incontravano nella programmazione di campioni ad hoc per ogni specifica indagine. In particolare la scarsa utilizzazione dello strumento informatico non consentiva di effettuare, in tempi brevi e con una certa sistematicità, le due operazioni connesse con il primo stadio di campionamento:

a) la stratificazione dei comuni con procedimenti variabili da indagine ad indagine;

b) la selezione dei comuni campione, con la possibilità di estrarre anche più di un comune per strato e utilizzando probabilità di selezione variabili.

Poiché le esigenze conoscitive delle nuove indagini sulla popolazione richiedevano campioni di dimensioni ridotte rispetto a quelle del campione delle forze di lavoro, per diminuire il numero dei comuni campione era necessario raggruppare gli strati già formati. Il collassamento degli strati si otteneva facendo saltare uno o più dei criteri utilizzati (attività prevalente, zona altimetrica o settore statistico).

Soltanto all'inizio degli anni '80, anche in conseguenza del sempre maggiore sviluppo che l'informatica ha avuto all'interno dell'Istituto, si è cominciato ad effettuare sperimentazioni sulla stratificazione e ad impiegare procedimenti diversi nella formazione degli strati (Napolitano, Russo e Zannella 1983, Zannella 1984, Russo 1984, 1985, 1986). Le innovazioni apportate hanno riguardato quasi tutte le fasi della stratificazione:

a) La scelta delle basi territoriali viene fatta dipendere dai livelli territoriali di riferimento delle stime (domini di studio), che possono variare da indagine ad indagine; così per l'indagine sulla salute la stratificazione è stata effettuata all'interno delle regioni mentre per quella sulle strutture familiari si è proceduto alla stratificazione entro le ripartizioni geografiche.

b) La determinazione dei comuni autorappresentativi viene effettuata sulla base di una soglia di popolazione variabile, determinata in modo da assicurare un numero minimo d'interviste per rilevatore.

c) Per la stratificazione dei comuni non autorappresentativi si è cominciato ad utilizzare la popolazione residente, in modo da tenere sotto controllo l'ammontare e la variabilità della popolazione negli strati.

Tutto ciò se da un lato ha consentito di predisporre piani di campionamento indipendenti da quello adottato per le forze di lavoro, e quindi più mirati verso le specifiche indagini, dall'altro non può costituire una soluzione soddisfacente dei problemi connessi con la stratificazione dei comuni.

Infatti, ancora oggi la stratificazione dei comuni richiede tempi abbastanza lunghi a causa della scarsa flessibilità delle procedure fin qui predisposte, che sono utilizzabili soltanto per quel particolare campionamento cui si riferiscono. Inoltre non è provato che le soluzioni date di volta in volta ai diversi problemi (scelta delle variabili di stratificazione, tecniche utilizzate per la formazione degli strati, ecc.) siano le più efficienti, né si hanno informazioni sull'efficienza e la praticabilità di altri procedimenti di stratificazione.

Appare evidente che questo secondo punto è strettamente dipendente dal precedente, in quanto sperimentazioni sui procedimenti di stratificazione, intesi a valutarne l'efficienza e la praticabilità, possono essere condotte soltanto se si dispone di una procedura di stratificazione flessibile ed informatizzata.

3. LA PROCEDURA GENERALIZZATA PER LA STRATIFICAZIONE DEI COMUNI

La procedura che è stata predisposta non deve essere intesa come un metodo standard di stratificazione dei comuni applicabile a qualsiasi indagine campionaria sulla popolazione, ma piuttosto come uno strumento flessibile che, in ogni fase del processo di stratificazione, consente all'utilizzatore di poter scegliere tra le diverse soluzioni proposte.

Le scelte che possono essere operate riguardano:

- i domini di studio e le basi territoriali di stratificazione;
- il criterio per l'individuazione dei comuni autorappresentativi;
- le variabili da utilizzare per la stratificazione dei comuni non autorappresentativi;
- la tecnica da adottare per il raggruppamento dei comuni non autorappresentativi in base alle variabili scelte;
- le variabili obiettivo ai fini della valutazione della bontà della stratificazione ottenuta.

La procedura prevede come input:

1) un file di dati comunali contenente per ogni comune i codici identificativi, la popolazione residente e i valori di un ampio insieme di variabili tra le quali scegliere quelle da utilizzare per la stratificazione;

2) un file con le informazioni relative ai domini territoriali entro i quali si deve procedere alla stratificazione, riportante per ciascun dominio: il codice identificativo, il numero dei comuni campione, il numero dei comuni da estrarre da ciascuno strato, la numerosità del campione in secondo stadio, il numero minimo e massimo di interviste che possono essere assegnate a ciascun rilevatore;

3) la data di riferimento dell'indagine espressa in giorno, mese ed anno.

La procedura nella sua configurazione attuale, utilizza una serie di programmi SAS che consentono:

- la stima per comune della popolazione residente alla data di riferimento dell'indagine;
- la determinazione dei comuni autorappresentativi in base al metodo selezionato;
- il calcolo del numero degli strati in cui devono essere raggruppati i comuni non autorappresentativi all'interno di ciascun dominio di studio;
- la stratificazione dei comuni non autorappresentativi mediante le variabili e la tecnica di formazione degli strati che sono stati prescelti;
- l'analisi della varianza condotta sulle variabili obiettivo all'interno di ciascun dominio territoriale;

- il calcolo della varianza campionaria della stima dell'ammontare totale di ciascuna variabile obiettivo per dominio territoriale, per un disegno campionario che prevede l'estrazione di un solo comune per strato con probabilità proporzionale all'ampiezza demografica e la ripartizione proporzionale tra gli strati della numerosità del campione di secondo stadio.

Per quanto concerne quest'ultimo punto è bene precisare che a regime la procedura consentirà il calcolo della varianza campionaria delle stime anche nel caso di disegni campionari che prevedono l'estrazione di due o più comuni per strato utilizzando diverse tecniche di selezione (Fabbris e Zannella, 1988).

L'output della procedura è costituito da:

1) il file dei comuni universo ordinato per dominio territoriale, strato e codice identificativo del comune, nel quale sono riportati in corrispondenza di ciascun comune la popolazione residente e i valori delle variabili obiettivo;

2) il file degli strati ordinato per dominio territoriale e codice di strato, contenente per ogni strato il numero dei comuni universo e campione e le statistiche descrittive sia delle variabili obiettivo che della popolazione residente.

3) il file con le varianze campionarie, gli errori standard assoluti e relativi delle stime per dominio di studio.

La flessibilità della procedura e la possibilità di ripetere il procedimento in tempi brevi permettono all'utilizzatore di sperimentare più metodi di stratificazione e di scegliere la soluzione più idonea dopo l'analisi dei diversi risultati conseguiti.

Prima di passare ad esaminare in modo più approfondito le singole parti in cui è stata articolata la procedura, occorre far presente che la sua configurazione attuale non è ancora soddisfacente sotto l'aspetto informatico. Infatti i programmi predisposti sono incompleti in quanto non ricoprono tutte le opzioni possibili, inoltre l'utilizzazione della procedura richiede il richiamo del programma che di volta in volta deve essere mandato in esecuzione e la modifica delle specifiche al suo interno sulla base delle scelte effettuate.

La procedura prevista a regime deve essere interattiva e deve consentire all'utente di operare le scelte su opportuni pannelli che visualizzano le alternative possibili (Fabbris e Zannella, op. cit.).

4. COSTRUZIONE ED AGGIORNAMENTO DELLA BASE DEI DATI COMUNALI

Il primo problema che è stato affrontato nella messa a punto della procedura è stato quello della costruzione e dell'aggiornamento di un file di dati comunali, quale base per la stratificazione stessa. Il file è costituito da tanti records quanti sono i comuni ed ognuno di essi contiene, oltre al nome del comune e al suo codice identificativo (codice di provincia e codice di comune all'interno della provincia), i valori delle variabili da utilizzare per la stratificazione.

L'insieme delle variabili contenute nel file dei dati di base costituisce *il potenziale di stratificazione*, in quanto i metodi previsti dalla procedura generalizzata sono basati sull'utilizzazione di una o più di queste variabili o di fattori ottenuti come loro combinazioni lineari.

Affinché la procedura risulti efficiente per qualsiasi indagine campionaria sulla popolazione, è necessario che il *potenziale di stratificazione* assicuri una buona rappresentazione dei comuni rispetto alle caratteristiche che devono essere stimate con l'indagine stessa. Poiché le indagini campionarie che dovranno utilizzare questa procedura hanno, essenzialmente, lo scopo di fornire delle stime relative alle caratteristiche demografiche, economiche e sociali della popolazione, si è ritenuto che una buona base di dati possa essere formata da:

a) caratteri qualitativi relativi alla localizzazione geografica (ripartizione, regione, provincia, unità sanitaria locale) e alla caratterizzazione del territorio (settore statistico e zona altimetrica), che, oltre a costituire in molti casi delle buone variabili di stratificazione spesso vengono utilizzate per definire i domini di studio o le basi territoriali di stratificazione;

b) distribuzione della popolazione residente secondo le modalità dei caratteri socio-demografici rilevati nell'ultimo censimento della popolazione (sesso, classi di età, stato civile, titolo di studio, ramo di attività economica, ecc.);

c) variabili relative alle famiglie e alle abitazioni, rilevate anch'esse in occasione del censimento della popolazione (famiglie secondo la tipologia, abitazioni per titolo di godimento e servizi installati, ecc.);

d) variabili relative al movimento anagrafico dei comuni (nascite, morti, iscrizioni e cancellazioni, popolazione residente).

La costruzione del file dei dati di base ha richiesto in primo luogo la predisposizione del file dei dati di censimento mediante l'estrazione delle variabili d'interesse dall'archivio dei dati comunali di censimento. Questa operazione è stata effettuata utilizzando un'apposita procedura per la gestione dell'archivio messa a punto nell'ambito dei lavori della Commissione (Gaggiotti e Zucchegna, 1985).

È stato, quindi, creato un file dei dati anagrafici per ognuno degli anni che vanno dal 1981 al 1985, ultimo anno per il quale erano disponibili le informazioni sul movimento anagrafico al momento della progettazione della procedura.

I files anagrafici contengono le seguenti informazioni:

codice di ripartizione geografica	
codice di regione	
codice di provincia	
codice di comune	
nome del comune	
codice di avviamento postale	
popolazione residente	al 31 dicembre
popolazione maschile	» » »
numero di famiglie	» » »
componenti delle famiglie	» » »
numero di convivenze	» » »
membri permanenti delle convivenze	» » »
nati vivi nell'anno	
morti nell'anno	

Mediante un apposito programma si è proceduto alla fusione del file dei dati di censimento con quelli relativi al movimento anagrafico, dopo averli ordinati per codice di provincia e di comune, ottenendo un nuovo file contenente sia le variabili di censimento che quelle anagrafiche per gli anni dal 1981 al 1985.

Poiché tra un anno e il successivo le liste dei comuni possono non coincidere, o per la nascita di nuovi comuni in seguito alla scissione di uno o più comuni o per la soppressione di alcuni comuni assorbiti da altri, il programma forma anche la lista dei comuni nuovi e quella dei comuni soppressi.

Dovendo il file dei dati di base risultare costituito dai comuni esistenti alla data più recente di disponibilità dei dati, si è proceduto ad aggiornare sia i dati di censimento che quelli relativi al movimento anagrafico per gli anni dal1981 al 1984 sulla base della situazione al 31 dicembre 1985, utilizzando le informazioni sulle variazioni territoriali intervenute in tale periodo (Istat, 1986). Dopo l'aggiornamento si è passati da 8086 comuni a 8090, in quanto dalla data del censimento sono stati costituiti 4 nuovi comuni:

Comune di nuova formazione			Comune dal quale si è formato		
PROV	COM	NOME	PROV	COM	NOME
83	108	Torrenova	83	079	San Marco
90	087	S. Maria Chochinas	90	079	Valledoria
91	103	Cardedu	91	026	Cairo
92	105	Quartuccio	92	009	Cagliari

L'aggiornamento ha anche comportato il calcolo dei valori delle variabili di censimento e di quelle anagrafiche per gli anni 1981-84 sia per i comuni nuovi che per quelli da cui questi sono derivati. La ricostruzione è stata effettuata ripartendo i valori delle variabili proporzionalmente alla popolazione residente nei due comuni al 31/12/1985.

Utilizzando i programmi predisposti sarà possibile procedere all'aggiornamento immediato del file dei dati di base al 31 dicembre di ciascun anno successivo non appena disponibili i dati del movimento anagrafico dei comuni.

5. LA STIMA DELLA POPOLAZIONE RESIDENTE ALLA DATA DI RIFERIMENTO DI UN'INDAGINE CAMPIONARIA

Nelle indagini campionarie sulla popolazione la conoscenza dei residenti in ciascun comune alla data di riferimento dell'indagine è necessaria per:

a) la stratificazione dei comuni;

b) la determinazione della numerosità campionaria nei singoli strati;

c) il calcolo delle probabilità di estrazione dei comuni nel caso di selezione con probabilità proporzionale all'ampiezza;

d) il calcolo dei coefficienti di espansione.

Tenendo presente che la programmazione del piano di campionamento deve precedere di qualche mese (da tre a sei) il periodo di esecuzione dell'indagine e che i dati relativi al movimento anagrafico dei comuni sono disponibili su supporto informatico con uno sfasamento di 6-8 mesi, si è predisposto un programma per la stima a breve termine (9-14 mesi) della popolazione residente in ciascun comune.

Nella messa a punto del programma si è dovuto tener conto che i dati sul movimento anagrafico fanno riferimento alla popolazione complessiva dei comuni (famiglie e convivenze) mentre nella quasi totalità delle indagini campionarie dell'Istat la popolazione di riferimento è costituita dai soli componenti le famiglie e che dati attendibili sulla popolazione residente per comune distinta tra componenti le famiglie e membri delle convivenze si hanno soltanto in occasione dei censimenti della popolazione.

Pertanto si è proceduto dapprima alla stima della popolazione residente e quindi alla sua ripartizione tra famiglie e convivenze sulla base delle percentuali riscontrate al censimento.

La stima della popolazione residente alla data di riferimento dell'indagine è stata ricavata mediante estrapolazione di una funzione lineare interpolante la serie storica della popolazione residente al 1° gennaio di ciascun anno. Poiché i valori utilizzabili vanno dal 1982 all'ultimo anno disponibile, nella predisposizione della metodologia e del relativo programma si è dovuto stabilire se tener conto di tutta la serie storica o solo degli anni più recenti.

Si è ritenuto opportuno, prima di procedere alla messa a punto del programma definitivo di effettuare alcune sperimentazioni e a tale scopo è stato predisposto un programma che prevedendo come input la sola data di riferimento dell'indagine (espressa in giorno, mese ed anno) consente il calcolo:

a) del tempo espresso in anni e frazioni di anno intercorrente tra la data dell'ultimo dato disponibile sulla popolazione residente e quella di riferimento dell'indagine;

b) dei parametri delle rette che, in ciascun comune interpolano la serie storica costituita dagli ultimi m anni, con $m = 2, 3, \ldots n$ dove n è il numero totale di anni disponibili);

c) della stima della popolazione residente in ciascun comune alla data di riferimento dell'indagine.

La sperimentazione è consistita nello stimare per ciascun comune la popolazione al 1° gennaio 1986 utilizzando una prima volta i dati relativi agli ultimi due anni (m = 2), una seconda volta quelli relativi agli ultimi tre anni (m = 3) e quindi quelli relativi agli ultimi 4 anni (m = 4) e nel confrontare i tre indici che misurano l'accostamento tra la popolazione stimata in ciascun comune e quella effettiva.

Poiché i risultati della sperimentazione non hanno evidenziato differenze apprezzabili tra i tre indici, anche se una leggera preferenza può essere espressa per m = 3, il programma definitivo effettua la stima della popolazione residente mediante estrapolazione di una serie i cui parametri vengono determinati interpolando con il metodo dei minimi quadrati i valori della popolazione residente relativi agli ultimi tre anni disponibili.

Prima di chiudere questo paragrafo è opportuno ricordare che attualmente il calcolo delle probabilità di selezione dei comuni e la ripartizione fra gli strati del campione di secondo stadio vengono effettuati sulla base della popolazione residente relativa all'ultimo anno di disponibilità dei dati sul movimento anagrafico, mentre i coefficienti di espansione vengono determinati sulla base della popolazione stimata alla data di riferimento dell'indagine, e ciò può comportare un aumento della varianza campionaria delle stime.

6. I DOMINI DI STUDIO E LE BASI TERRITORIALI DI STRATIFICAZIONE

Come è stato già detto un dominio di studio è un sottogruppo della popolazione per il quale si desiderano ottenere stime con una prefissata precisione. Ai fini della programmazione del campione e quindi della stratificazione è utile distinguere tra due tipi di domini di studio (Verma, 1982):

1. domini territoriali (geographic classes), costituiti dalle popolazioni residenti in particolari raggruppamenti di comuni (ripartizioni geografiche, regioni, province, unità sanitarie locali, ecc.), che danno luogo a subpopolazioni disgiunte rispetto sia alle unità di primo che di secondo stadio;

2. sottoclassi (cross-classes), ottenute suddividendo la popolazione in base a determinate caratteristiche socio-demografiche od economiche (classi di età, titolo di studio, stato civile, condizione professionale, ecc.); poiché la popolazione di uno stesso comune si può distribuire su più sottoclassi, queste formano delle partizioni sovrapposte rispetto alle unità di primo stadio.

È evidente che sia il calcolo della numerosità che la stratificazione e la selezione delle unità di primo stadio possono essere pianificate a livello dei singoli domini soltanto nel caso in cui questi sono costituiti da domini territoriali, pertanto nel seguito con il termine di «domini di studio» si farà riferimento esclusivamente a quest'ultima accezione.

Le basi territoriali di stratificazione costituiscono delle suddivisioni del territorio finalizzate ad esigenze organizzative ed amministrative, che possono coincidere, ma non necessariamente, con i domini di studio, e vanno disegnate in modo che ciascuna di esse sia completamente distribuita su un unico dominio.

Gli strati sono raggruppamenti di comuni, appartenenti alla stessa base territoriale, finalizzati al miglioramento dell'efficienza del disegno campionario; devono, quindi, essere formati in modo da risultare al loro interno il più possibile omogenei rispetto alle variabili oggetto d'indagine.

La stratificazione che si ottiene operando all'interno delle basi territoriali è in genere meno efficiente di quella che si otterrebbe, con lo stesso metodo e a parità di numero di strati, raggruppando i comuni all'interno dell'intero dominio. Infatti in questo secondo caso è possibile inserire nello stesso strato comuni simili anche se appartenenti a basi territoriali diverse, ottenendo così strati più omogenei.

La divisione dei domini di studio in aree che possono presentare una scarsa omogeneità delle variabili d'interesse va, quindi, effettuata soltanto se motivata da effettive esigenze organizzative. Così la suddivisione del territorio provinciale in settori statistici non è indispensabile ai fini del miglioramento dell'organizzazione del lavoro sul campo, e pertanto va mantenuta soltanto se trova giustificazione in termini di maggiore efficienza rispetto ad altri criteri di raggruppamento dei comuni.

Attualmente la procedura richiede la predisposizione di un file contenente le informazioni relative a ciascun dominio territoriale, mentre nella versione generalizzata esso verrà generato dal programma che calcola le numerosità campionarie per i diversi stadi di campionamento in base ai costi e ai livelli attesi di attendibilità delle stime.

I domini territoriali previsti dalla procedura possono essere ripartizioni, regioni, province, USL, raggruppamenti di comuni per classi di ampiezza demografica o singoli comuni, e vengono definiti su indicazione del reparto dell'Istat responsabile dell'indagine.

Il file è formato da tanti records quanto sono i domini territoriali considerati e contiene le seguenti variabili:

- codice di dominio territoriale
- numero di comuni campione
- numero di comuni non autorappresentativi da estrarre da ciascuno strato
- tasso di campionamento finale
- numero minimo di interviste per rilevatore

Per quanto riguarda il numero di comuni da estrarre da ciascuno strato la scelta più conveniente, almeno dal punto di vista dell'efficienza della stratificazione, è quella di selezionare un solo comune per strato, poiché la varianza della stima si riduce con l'aumentare del numero degli strati.

Tuttavia la riduzione è poco sensibile nel caso in cui le variabili oggetto di indagine sono scarsamente correlate con quelle utilizzate per la stratificazione (Cochran 1977, Kish e Anderson 1978). Pertanto nelle indagini campionarie in cui vengono rilevate numerose variabili tutte ugualmente importanti, la decisione di estrarre un solo comune per strato non può essere completamente giustificata dall'aumento della precisione delle stime conseguente ad una stratificazione più fine. Inoltre, la scelta di un solo comune per strato non consente una stima corretta della varianza campionaria, in quanto per stimare la varianza è necessario collassare gli strati e ciò comporta una distorsione positiva (Hansen, Hurwitz and Madow, 1953).

Per una trattazione più completa del problema si rimanda al lavoro di Fabbris (1988) in cui sono riportati i risultati e le conclusioni cui è pervenuto dopo numerose sperimentazioni; qui interessa sol-

tanto evidenziare che il numero dei comuni da selezionare da ciascuno strato può variare da indagine ad indagine e per la stessa indagine anche tra i diversi domini territoriali.

Un'ultima considerazione va svolta sulla ripartizione della numerosità del campione di secondo stadio tra gli strati appartenenti allo stesso dominio territoriale.

La teoria suggerisce di utilizzare un tasso di campionamento finale variabile da strato a strato in modo da massimizzare l'efficienza per unità di costo, ma una tale soluzione risulta di limitata importanza per le indagini campionarie sulla popolazione che vengono usualmente effettuate all'Istat. Infatti, le rilevazioni, come più volte detto, riguardano numerose variabili ed un incremento dell'efficienza per una di esse può comportare una riduzione per un'altra. Inoltre, in molti casi il costo e la variabilità non variano molto da strato a strato e, anche quando ciò si verifica, non sempre le informazioni disponibili prima dell'indagine consentono la programmazione di un campione ottimo. Per ultimo, la variabilità del tasso di campionamento comporta alcune complessità aggiuntive nell'elaborazione dei risultati.

Per tutte queste ragioni si è ritenuto più idonea e praticabile l'adozione di una ripartizione proporzionale e quindi di un tasso di campionamento uguale per tutti gli strati all'interno dello stesso dominio territoriale.

7. LA DETERMINAZIONE DEI COMUNI AUTORAPPRESENTATIVI E DEL NUMERO DEGLI STRATI NON AUTORAPPRESENTATIVI

Una particolare tecnica di stratificazione è quella basata sulla suddivisione, all'interno di ciascun dominio territoriale, dei comuni in due gruppi: il primo costituito dai comuni autorappresentativi e il secondo da tutti gli altri comuni.

Ciascun comune autorappresentativo forma uno strato a se stante all'interno del quale si procede ad un campionamento semplice delle unità finali di rilevazione. Viene chiamato autorappresentativo in quanto l'intera popolazione del comune viene ad essere rappresentata da un campione casuale della popolazione stessa.

Per i comuni del secondo gruppo si procede invece al campionamento a due stadi, per cui la popolazione di un comune può essere rappresentata da un campione di persone estratto da un altro comune dello stesso strato.

Questa tecnica produce un alto guadagno nella precisione delle stime quando si ha un'elevata asimmetria nella distribuzione della popolazione, ed è pertanto particolarmente utile nella situazione italiana, in cui ci sono pochi comuni di grandi dimensioni ed un numero elevato di comuni di piccola ampiezza demografica.

Il problema da risolvere è quello della determinazione, all'interno di ciascun dominio territoriale, della soglia di popolazione oltre la quale un comune è autorappresentativo in modo da massimizzare l'efficienza degli stimatori.

Soluzioni approssimate sono state date da numerosi autori (Dalenius 1950, Glasser 1962, Hidiroglou 1977) nel caso in cui le unità non autorappresentative sono estratte mediante un campionamento casuale semplice senza reimmissione.

Più recentemente Hidiroglou (1986) ha sviluppato un algoritmo per determinare il valore esatto della soglia in modo da minimizzare la numerosità campionaria avendo fissato il livello di precisione della stima.

Le diverse soluzioni proposte non risultano tuttavia direttamente applicabili al disegno campionario che viene considerato in questo lavoro, che prevede la stratificazione dei comuni non autorappresentativi e la possibilità di una loro selezione con probabilità variabile; pertanto la determinazione della soglia ottima per questo specifico piano di campionamento resta ancora un problema aperto.

L'ampiezza demografica minima dei comuni autorappresentativi può anche essere determinata in base a criteri operativi, oltre che teorici. Così con riferimento ad un generico dominio territoriale, se si indica con:

r = numero di rilevatori
k = numero minimo di interviste da assegnare a ciascun rilevatore
f = tasso di campionamento finale
m= numero medio di componenti per famiglia

un comune per essere autorappresentativo deve avere una popolazione non minore della soglia A, dove:

$$A = r \cdot k \cdot m / f \qquad (1)$$

in modo da avere nel comune un campione di almeno $k \cdot r$ famiglie, così da assicurare un minimo di k interviste per rilevatore.

Ad esempio per un dominio territoriale con un tasso di campionamento finale f = 1/500 e con un numero medio di componenti per famiglia m = 3, per il quale è previsto un numero minimo di k = 15 famiglie da intervistare per rilevatore, nel caso in cui venga impiegato un solo rilevatore si ha che la soglia A è uguale:

15 x 3 x 500 = 22.500.

Nel prospetto che segue sono riportate le ampiezze demografiche minime per diversi valori di k e di f nel caso in cui si voglia utilizzare un solo rilevatore per comune.

Prospetto 1 - Soglia per la determinazione dei comuni autorappresentativi in funzione del tasso di campionamento e del numero di interviste per rilevatore

TASSO DI CAMPIONAMENTO (f)	N. di interviste per rilevatore		
	10	15	20
1/100	3.000	4.500	6.000
1/200	6.000	9.000	12.000
1/300	9.000	13.500	18.000
1/400	12.000	18.000	24.000
1/500	15.000	22.500	30.000
1/600	18.000	27.000	36.000
1/700	21.000	31.500	42.000
1/800	24.000	36.000	48.000
1/900	27.000	40.500	54.000
1/1.000	30.000	45.000	60.000

È importante osservare che l'ampiezza demografica minima non deve essere necessariamente la stessa per tutti i domini territoriali. Infatti nel caso in cui si adotti la «soglia ottima» il valore calcolato dipende dalla precisione desiderata, dalla variabilità del carattere e dalla distribuzione della popolazione nei comuni, quantità che possono risultare variabili da dominio a dominio. Così come possono

essere diversi il tasso di campionamento e il numero di interviste per rilevatore, da cui dipende il valore della soglia quando si deve utilizzare il criterio imposto dalle esigenze pratiche.

Il programma che è stato messo a punto per la determinazione dei comuni autorappresentativi prevede come input il file dei dati di base e quello contenente le informazioni relative ai domini territoriali e utilizza la (1) per il calcolo del valore della soglia.

Nel file dei dati di base viene introdotta da programma una nuova variabile che assume il valore 1 se il comune è autorappresentativo, ossia se il comune ha una popolazione residente maggiore od uguale al valore della soglia calcolato per il dominio di appartenenza, e il valore 2 nel caso contrario. Quindi vengono creati due nuovi files separando i comuni autorappresentativi dagli altri comuni.

Il programma provvede anche all'inserimento nel file dei domini territoriali delle seguenti variabili:

a) numero dei comuni autorappresentativi;

b) numero dei comuni non autorappresentativi da campionare, ottenuto come differenza tra il numero totale dei comuni campione e il numero dei comuni autorappresentativi;

c) numero degli strati in cui devono essere raggruppati i comuni non autorappresentativi ottenuto come rapporto tra la numerosità del campione dei comuni non auorappresentativi e il numero dei comuni che devono essere selezionati da ciascuno strato.

8. LA SCELTA DELLE VARIABILI DI STRATIFICAZIONE

La teoria del campionamento stratificato, così come molti altri importanti aspetti del campionamento da popolazioni finite, ha avuto uno sviluppo soddisfacente con riferimento al problema della stima di un parametro relativo ad una sola variabile; nella pratica, invece, le indagini campionarie sulla popolazione vengono effettuate per raccogliere una molteplicità d'informazioni, per cui, si può verificare che un campione «ottimo» per una determinata variabile presenti un'efficienza ridotta per altre.

Per dare una corretta soluzione ai problemi che si presentano nella determinazione della stratificazione «ottima» è necessario, quin-

di, che per ogni indagine venga identificata la variabile «guida», o, nel caso di più variabili, venga definito un sistema di ponderazione che consenta comunque di riportarsi al caso di una sola variabile. Una volta determinata la stratificazione «ottima» per questa variabile è possibile calcolare la perdita di efficienza per ogni altra variabile oggetto di rilevazione, purché si abbiano le necessarie informazioni.

Poiché la stratificazione «ottima» si ottiene suddividendo la popolazione in gruppi al loro interno più omogenei possibile rispetto alla variabile guida, è evidente che se per formare gli strati si utilizza un'altra variabile, questa risulterà tanto più idonea quanto più sarà correlata con la variabile guida.

Nel caso in cui si vuole utilizzare una sola variabile, la scelta è immediata se si dispone delle informazioni sulle correlazioni esistenti tra le variabili di stratificazione e la variabile guida.

Dovendo scegliere un gruppo di variabili da un insieme più ampio è necessario disporre non solo delle stime dei coefficienti di correlazione tra la variabile guida e le potenziali variabili di stratificazione, ma anche della stima della matrice delle correlazioni tra queste ultime. Infatti, il guadagno che si ricava nell'introdurre una nuova variabile nel procedimento di stratificazione dipende, oltre che dall'intensità del legame con la variabile guida, da quella dei legami con le altre variabili di stratificazione precedentemente considerate (Zannella 1983). È evidente che sono da preferire variabili incorrelate tra loro ma correlate con la variabile guida, poiché se due variabili di stratificazione sono perfettamente correlate usare l'una o l'altra dà lo stesso guadagno che usarle entrambe.

Da quanto detto si evince che le variabili di stratificazione possono risultare diverse da indagine ad indagine e perfino tra i diversi domini territoriali nell'ambito della stessa rilevazione.

Il problema della scelta delle variabili di stratificazione nelle indagini multiscopo si presenta, quindi, piuttosto complesso. Tuttavia, nella pratica non è molto vantaggioso spendere eccessivo lavoro per la ricerca di una soluzione ottimale, in quanto scelte alternative possono dare risultati ugualmente buoni. Hansen, Hurwitz e Madow (1953) riportano uno studio comparativo effettuato da Nisselson, Goodman e Berger su un campione di villaggi agricoli negli U.S.A. utilizzando due diversi metodi di stratificazione. Sebbene i fattori di

stratificazione fossero molto diversi, le varianze campionarie per i due approcci erano dello stesso ordine di grandezza, nessuno era uniformemente migliore dell'altro.

O' Muircheartaigh (1977) nell'esaminare i problemi connessi con la programmazione di piani di campionamento da utilizzare per indagini sulla fecondità, individua in una buona suddivisione geografica uno dei migliori criteri di stratificazione per i Paesi sviluppati, essendo questa classificazione correlata con molte variabili socioeconomiche. Inoltre, per i Paesi nei quali è stato effettuato il censimento consiglia di utilizzare i dati sulla densità della popolazione per una stratificazione più raffinata.

L'adozione di questo criterio per la stratificazione dei comuni italiani, con l'eventuale impiego dell'ampiezza demografica al posto della densità, va, comunque, preceduta da una serie di sperimentazioni.

Quando, come spesso accade, i caratteri di rilevazione sono numerosi, tutti ugualmente importanti e non sempre correlati tra di loro, non è possibile spiegarne la variabilità ricorrendo ad una sola o comunque a poche variabili di stratificazione. In questi casi è preferibile utilizzare più variabili in modo da coprire i diversi campi di interesse dell'indagine.

Si è quindi messo a punto un programma che consente la costruzione di un insieme di fattori ortogonali, ricavati mediante l'analisi in componenti principali di una serie di indicatori sociodemografici. Gli indicatori presi in considerazione sono stati scelti tra i 40 indicatori di sviluppo proposti dall'Istituto di Ricerca per lo Sviluppo Sociale delle Nazioni Unite (UNRISD), per i quali è risultato possibile il calcolo a livello comunale utilizzando le variabili di censimento e quelle anagrafiche contenute nel file dei dati di base.

Gli indicatori calcolati sono:

Z1 = densità
Z2 = rapporto di dipendenza
Z3 = indice di vecchiaia
Z4 = tasso di analfabetismo
Z5 = % pop. in età 6-13 anni iscritta alla scuola dell'obbligo
Z6 = % pop. in età 14-19 anni iscritta a corsi regolari
Z7 = % pop. in età 14-19 anni iscritta a corsi professionali
Z8 = % diplomati e laureati

Z9 = % popolazione attiva
Z10 = tasso di disoccupazione
Z11 = % addetti in agricoltura
Z12 = % addetti nell'industria
Z13 = % addetti in altre attività
Z14 = % imprenditori e liberi professionisti
Z15 = % lavoratori indipendenti e coadiuvanti
Z16 = % lavoratori dipendenti
Z17 = numero medio di componenti per famiglia
Z18 = numero medio di occupanti per stanza
Z19 = % abitazioni occupate in proprietà
Z20 = % ab. occupate sprovviste di acqua potabile e gabinetto
Z21 = % ab. occupate provviste di elettricità

Il programma utilizza la procedura SAS «PRINCOMP» che prevede come input il file dei dati di base e consente di condurre l'analisi in componenti principali utilizzando diverse opzioni.

Così l'analisi può essere fatta in corrispondenza di qualunque livello territoriale (Italia, ripartizioni, regioni, province, etc.), specificando il dominio entro il quale si vuole effettuare l'elaborazione e dopo aver ordinato i dati secondo il codice di dominio.

Gli autovalori vengono calcolati per default sulla matrice di correlazione, mentre mediante l'opzione «COV» il calcolo viene fatto sulla matrice delle varianze e covarianze.

Infine, le unità di osservazione (i comuni) vengono considerati per default con lo stesso peso, mentre mediante opportuna specificazione è possibile assegnare loro un peso proporzionale alla popolazione residente o a qualsiasi altra variabile.

Il programma fornisce come output:

- le statistiche descrittive degli indicatori Z1-Z21;
- gli autovalori e gli autovettori;
- i coefficienti di correlazione fra le variabili e i fattori;
- il file dei dati di base contenente per ciascun comune tutte le variabili originali e i valori delle componenti principali.

In definitiva, per quanto concerne la scelta delle variabili di stratificazione, la procedura prevede la possibilità di poter scegliere una

o più variabili tra quelle contenute nel file dei dati di base, comprese le componenti principali, e di effettuare più stratificazioni utilizzando di volta in volta variabili diverse. L'efficienza relativa delle singole stratificazioni può essere valutata con riferimento ad alcune variabili obiettivo scelte tra quelle presenti nel file dei dati di base, tenuto conto delle finalità conoscitive dell'indagine per la quale il campione viene programmato.

9. LA FORMAZIONE DEGLI STRATI

I metodi che la procedura mette a disposizione per la formazione degli strati sono diversi a seconda che si utilizzi una sola variabile o più variabili di stratificazione.

Quando viene impiegata una sola variabile gli strati vengono determinati mediante una sua opportuna suddivisione in classi. Una naturale estensione di questa procedura al caso di più variabili è quella di formare gli strati tramite una classificazione incrociata.

È evidente che questo metodo diventa impraticabile non appena il numero delle variabili è un poco elevato, a causa del gran numero di strati potenziali che possono essere prodotti dalla classificazione incrociata. Può così risultare che il numero degli strati formati sia superiore a quello desiderato e che molti di essi siano costituiti da una o poche unità di primo stadio.

Golden e Yeomans (1963) hanno fornito un'espressione che consente di calcolare il numero massimo di variabili che possono essere utilizzate per una classificazione incrociata, in funzione del numero dei comuni che devono essere raggruppati, del numero delle classi in cui viene suddivisa ciascuna variabile e in modo da assicurare un numero minimo di unità per strato.

Indicando con N il numero di comuni non autorappresentativi di un generico dominio territoriale, con P il numero minimo di comuni che devono essere contenuti in ogni strato e con m il numero di classi in cui è suddivisa ciascuna variabile di stratificazione, il numero massimo di variabili che è possibile utilizzare è dato da:

$$s(max) = INT\ (\log(N/P)/\log(m)) \qquad (9.1)$$

Il prospetto che segue evidenzia le limitazioni sul numero delle variabili di stratificazione in corrispondenza a diversi valori di N, P e m.

Prospetto 2 - Numero massimo di variabili di stratificazione per diverse combinazioni di N P e m.

N	P=2				P=5				P=10			
	m=2	m=3	m=4	m=5	m=2	m=3	m=4	m=5	m=2	m=3	m=4	m=5
20	3	2	1	1	2	1	1	0	1	0	0	0
40	4	2	2	1	3	1	1	1	2	1	1	0
60	4	3	2	2	3	2	1	1	2	1	1	1
80	5	3	2	2	4	2	2	1	3	1	1	1
100	5	3	2	2	4	2	2	1	3	2	1	1
200	6	4	3	2	5	3	2	2	4	2	2	1
400	7	4	3	3	6	3	3	2	5	3	2	2
600	8	5	4	3	6	4	3	2	5	3	2	2
800	8	5	4	3	7	4	3	3	6	3	3	2
1000	8	5	4	3	7	4	3	3	6	4	3	2

Tenuto conto di queste limitazioni, per la stratificazione basata su due o più variabili è prevista l'utilizzazione della cluster analysis che consente di formare un numero qualsiasi di strati indipendentemente dal numeri di variabili scelte.

Di seguito vengone elencati i metodi che possono essere richiamati dalla procedura, distinti a seconda del numero di variabili che vengono impiegate:

A. Metodi che utilizzano una sola variabile.

A1. Uguale numero di comuni per strato;

A2. Uguaglianza dell'ammontare totale della variabile di stratificazione in ciascuno strato, metodo suggerito da Mahalanobis (1952) e Hansen, Hurwitz e Madow (1953);

A3. Procedimento di Dalenius e Hodges (1958), che si basa sulla costruzione di intervalli uguali sulla cumulata della radice quadrata della frequenza della variabile di stratificazione;

A4. Minimizzazione della variabilità all'interno degli strati mediante cluster analysis della variabile di stratificazione;

A5. Inputazione degli estremi delle classi.

B. Metodi che utilizzano due o più variabili.

B1. Classificazioni incrociate mediante imputazione degli estremi delle classi di ciascuna variabile di stratificazione;

B2. Determinazione della modalità prevalente, quando le variabili di stratificazione si riferiscono alle percentuali di popolazione relative alle modalità di un carattere qualitativo e alle classi di un carattere quantitativo;

B3. Cluster analysis delle variabili di stratificazione mediante il metodo delle k medie.

Per ogni tecnica di formazione degli strati è stato predisposto uno specifico programma e a regime la scelta potrà essere effettuata mediante selezione da effettuare su un pannello contenente l'elenco dei diversi metodi.

Il programma richiede l'immissione da tastiera delle variabili da utilizzare per la formazione degli strati, i cui valori vengono letti dal file dei dati di base, mentre il numero degli strati che devono essere formati in ciascun dominio viene letto dal file dei domini territoriali.

Ciascun programma prevede come output:

a) il file dei dati di base dei comuni non autorappresentativi in cui viene inserita una nuova variabile corrispondente al codice di strato;

b) il file degli strati ordinati per dominio territoriale e codice di strato, contenente per ciascuno strato il numero dei comuni universo e campione, la popolazione totale, media, minima e massima, e il numero di unità di secondo stadio da campionare;

c) i risultati dell'analisi della varianza condotta per ognuna delle variabili obiettivo all'interno di ciascun dominio territoriale.

Infine un ultimo programma fonde il file del punto a) con quello dei comuni autorappresentativi ed assegna un numero progressivo di strato su 6 cifre: le prime due indicano il dominio, la terza il tipo di comune e le ultime 3 il numero d'ordine progressivo dello strato all'interno del dominio e del tipo di comune.

È prevista la stampa del file finale ordinato per dominio territoriale, strato e popolazione residente. Per ogni comune vengono stampati:

- codice del dominio territoriale
- codice dello strato
- codice della provincia
- codice e nome del comune
- codice di avviamento postale
- popolazione residente alla data di riferimento dell'indagine
- valori della variabile obiettivo

Quest'ultimo file è stato predisposto in modo da essere utilizzato come input nella procedura per la selezione dei comuni campione.

10. FINALITÀ E MODALITÀ DELLA SPERIMENTAZIONE

10.1 I procedimenti di stratificazione

La sperimentazione effettuata ha avuto lo scopo di verificare la procedura proposta e i programmi SAS predisposti e di mettere a confronto diversi procedimenti di stratificazione per valutarne sia l'efficienza che la praticabilità. Essa è stata condotta su tre regioni (Piemonte, Toscana e Calabria), considerate ciascuna come un dominio di studio, per cui i confronti fra le diverse stratificazioni sono stati effettuati comparando le varianze campionarie delle stime regionali.

All'interno di ciascuna regione sono state prese in considerazione tre diverse basi territoriali:

a) le province

b) i settori statistici

c) le zone altimetriche presenti all'interno di ciascuna provincia.

Per la stratificazione dei comuni non autorappresentativi sono state sperimentate le seguenti variabili:

a) popolazione attiva in condizione professionale per settore di attività economica (agricoltura, industria, altre attività);

b) popolazione residente;

c) prime due componenti principali risultanti dall'analisi condotta a livello provinciale sui 21 indicatori socio-demografici.

Per la formazione degli strati sono stati utilizzati la cluster analysis, il criterio della modalità prevalente e quello basato sull'uguale ammontare del carattere negli strati. Nel complesso sono state sperimentate le seguenti stratificazioni:

1. Variabili di stratificazione: popolazione in condizione professionale per ramo di attività economica

AE1 = stratificazione per provincia, settore statistico, zona altimetrica ed attività economica prevalente;

AE2 = cluster analysis delle % di popolazione per ramo di attività economica all'interno di ciascuna provincia;

AE3 = cluster analysis delle % di popolazione per ramo di attività economica per provincia e zona altimetrica;

AE4 = cluster analysis delle % di popolazione per ramo di attività economica, per provincia e settore statistico.

2. Variabile di stratificazione: popolazione residente

POP2 = classi di popolazione con uguale numero di comuni all'interno di ciascuno strato;

POP2 = classi di popolazione con uguale ammontare della popolazione all'interno di ciascuno strato;

POP3 = cluster analysis della popolazione residente all'interno di ciascuna provincia;

POP4 = cluster analysis della popolazione residente per provincia e zona altimetrica;

POP5 = cluster analysis della popolazione residente per provincia e settore statistico.

3. Variabili di stratificazione: prime due componenti principali

PR1 = cluster analysis delle prime due componenti principali all'interno di ciascuna provincia;

PR2 = cluster analysis delle prime due componenti principali per provincia e zona altimetrica;

PR3 = cluster analysis delle prime due componenti principali per provincia e settore statistico.

10.2 Il disegno campionario e le dimensioni del campione

Per valutare l'efficienza delle diverse stratificazioni prese in esame si è preso in considerazione un disegno campionario che prevede:

a) l'estrazione di un solo comune per strato con probabilità proporzionale all'ampiezza demografica;

b) l'adozione di un tasso di campionamento costante per ogni strato e l'estrazione delle unità di secondo stadio con uguale probabilità e senza reimmissione.

La scelta di questo piano di campionamento, e in particolare quella della selezione di un solo comune per strato, è motivata dal fatto che essa consente una stratificazione più fine e quindi permette di valutare il «guadagno» massimo che si può conseguire con ciascuna stratificazione.

Per il calcolo della varianza campionaria si è fatto riferimento alla stima dell'ammontare totale di un carattere e allo stimatore proposto da Hansen e Hurwitz. L'espressione utilizzata per il calcolo della varianza è un'estensione al campionamento stratificato di quella riportata dal Cochran per un campionamento senza stratificazione (Cochran 1977, p. 295):

$$V(\hat{Y}) = \sum_{h=1}^{L} M_h^2 \sum_{i=1}^{N_h} \frac{M_{hi}}{M_h} (\bar{Y}_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^{L} M_h^2 \sum_{i=1}^{N_h} \frac{M_{hi} - m_h}{M_h} \frac{S_{2hi}^2}{m_h} \quad (2)$$

dove:

$\hat{Y}$ = stima dell'ammontare totale;
h = indice di strato;
N_h = numero dei comuni nello strato h
M_h = popolazione residente nello strato h
$\bar{Y}_h$ = media del carattere Y nello strato h
i = indice di comune
M_{hi} = popolazione residente nel comune hi
$\bar{Y}_{hi}$ = media del carattere Y nel comune hi

S^2_{2hi} = varianza tra le unità di secondo stadio nel comune hi

m_h = numerosità del campione in secondo stadio nello strato h.

I due termini a secondo membro della (2) rappresentano le componenti della varianza dovute rispettivamente al primo (V1) e al secondo (V2) stadio di campionamento.

Si verifica immediatamente che la (2) rimane valida anche quando l'indice h si riferisce ad uno strato costituito da un comune autorappresentativo; in questo caso la varianza di primo stadio imputabile a quello strato è uguale a zero.

La componente di primo stadio è funzione del numero dei comuni campione, che nel caso considerato coincide con il numero degli strati, e dalla variabilità tra i comuni all'interno degli strati, che a sua volta dipende dal procedimento di stratificazione adottato.

Per quanto riguarda la varianza di secondo stadio essa può essere scomposta nella differenza tra due quantità:

$$V2 = V2A - V2B \tag{3}$$

dove:

$$V2A = \frac{1}{f} \sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{hi} S^2_{2hi} \tag{4}$$

e

$$V2B = \sum_{h=1}^{L} M_h \sum_{i=1}^{N_h} S^2_{2hi} \tag{5}$$

La quantità V2A risulta indipendente dalla stratificazione, in quanto è funzione del tasso di campionamento e della somma delle devianze entro i comuni, e rappresenta la parte di varianza imputabile al secondo stadio di campionamento in assenza di stratificazione.

La quantità V2B è sempre indipendente dal tasso di campionamento, mentre risulta influenzata dalla stratificazione nel caso in cui l'ammontare della popolazione residente è variabile da strato a strato. Comunque, in genere, essa assume valori trascurabili e poco variabili con il procedimento di stratificazione.

Da quanto detto si evince che affinché un cambiamento di stratificazione comporti una riduzione rilevante nella varianza dello stimatore è necessario che si verifichino due condizioni:

1. la varianza di primo stadio deve costituire una componente importante della varianza complessiva;

2. la stratificazione deve comportare una riduzione sensibile della varianza di primo stadio.

La prima condizione dipende:

- dal modo in cui la variabilità del carattere in esame si scompone entro e tra i comuni;
- dal numero dei comuni autorappresentativi e dalla quota di popolazione in essi residente;
- dalla numerosità del campione in primo stadio;
- dal tasso di campionamento f.

Per tener conto delle diverse situazioni che si possono verificare nella pratica, sono state considerate diverse ipotesi circa la numerosità del campione in primo e in secondo stadio:

Per il primo stadio di campionamento:

1A. numero di comuni, e quindi di strati, uguale a quello del campione base delle forze di lavoro;

1B. numero di comuni, e quindi di strati, ridotto di circa la metà rispetto al campione base delle forze di lavoro.

Per il secondo stadio di campionamento:

2A. tasso di campionamento f = 1/500 corrispondente a quello usualmente utilizzato dall'Istat per indagini di medie dimensioni;

2B. tasso di campionamento f = 1/200 corrispondente al tasso medio utilizzato dal campione base delle forze di lavoro.

Poiché scopo della sperimentazione è essenzialmente quello di confrontare i diversi criteri di stratificazione dei comuni non autorappresentativi, al fine di non appesantire troppo le elaborazioni si è proceduto alla determinazione dei comuni autorappresentativi utilizzando una soglia costante uguale a 20.000 abitanti. Tale soglia assicura un numero minimo d'interviste per rilevatore pari a 33 per f = 1/200 e a 13 per f = 1/500.

Nel prospetto 3 è riportata la distribuzione del numero dei comuni autorappresentativi e del numero degli strati in cui devono essere raggruppati i comuni non autorappresentativi, per ciascuna provincia delle tre regioni considerate.

Prospetto 3 - Numero di strati per tipo di comune

PROVINCE	Comuni autorap.	Comuni non autorap.		Totale comuni	
		base	ridotto	base	ridotto
		PIEMONTE			
1. Torino	13	28	13	41	26
2. Vercelli	2	10	6	12	8
3. Novara	3	11	7	14	10
4. Cuneo	5	26	9	31	14
5. Asti	1	7	4	8	5
6. Alessandria	6	16	9	22	15
Piemonte	**30**	**98**	**48**	**128**	**78**
		TOSCANA			
45. Massa	2	5	2	7	4
46. Lucca	5	5	2	10	7
47. Pistoia	3	4	2	7	5
48. Firenze	9	14	3	23	12
49. Livorno	4	3	2	7	6
50. Pisa	5	7	2	12	7
51. Arezzo	3	4	2	7	5
52. Siena	2	11	4	13	6
53. Grosseto	2	9	3	11	5
Toscana	**35**	**62**	**22**	**97**	**57**
		CALABRIA			
78. Cosenza	7	18	8	25	15
79. Catanzaro	4	31	12	35	16
80. Reggio Calabria	1	21	8	22	9
Calabria	**12**	**70**	**28**	**82**	**40**

10.3 Le elaborazioni effettuate

Le elaborazioni hanno interessato la stima della frequenza assoluta delle persone che presentano le seguenti caratteristiche:

Y1 = laureati
Y2 = forniti di licenza di scuola media inferiore
Y3 = occupati
Y4 = disoccupati

Y5 = in cerca di prima occupazione
Y6 = in condizione professionale in agricoltura
Y7 = in condizione professionale nell'industria
Y8 = in condizione professionale nelle altre attività
Y9 = imprenditori e liberi professionisti
Y10 = lavoratori dipendenti

Le variabili considerate sono del tipo comunemente rilevato nelle indagini campionarie sulla popolazione effettuate dall'Istat. Come si evince dai valori riportati nel prospetto 4 ve ne sono alcune che misurano fenomeni poco frequenti con percentuali inferiori al 7% (Y1, Y4, Y5 e Y9) ed altre con un peso nella popolazione molto più rilevante: dal 17% a circa il 50%. Notevole è anche la diversità dei valori della variabilità tra i comuni, con valori del coefficiente di variazione che vanno dal 6 al 140%.

Anche per quanto riguarda le correlazioni, le variabili prescelte presentano una vasta gamma di possibilità: vi sono variabili fortemente correlate (valore assoluto di $r > 0.5$), altre che presentano valori di r moderati o addirittura prossimi allo zero (prospetto 5).

La procedura ha fornito questo output:

1. il file dei dati comunali contenente per ciascun comune il codice identificativo, il codice di dominio territoriale, la popolazione residente, i valori delle dieci variabili obiettivo e i codici di strato relativi alle dodici stratificazioni sperimentate;

2. il file degli strati, uno per ciascuna stratificazione, nei quali, in corrispondenza di ogni strato, sono riportati il numero dei comuni universo, la popolazione residente (totale, media, minima e massima) e alcune statistiche descrittive relative alle variabili obiettivo.

3. il file con le varianze campionarie (totali, di primo e di secondo stadio) degli stimatori dei totali delle variabili obiettivo per ognuna delle dodici stratificazioni.

Prospetto 4 - Statistiche descrittive delle variabili obiettivo

VARIABILI OBIETTIVO	media	min	max	std	cv
		PIEMONTE			
Y1 laurea	2,27	0,00	8,56	1,32	58,0
Y2 lic. scuola media inferiore	23,85	5,83	43,33	3,64	15,3
Y3 occupati	39,70	20,56	62,75	2,46	6,2
Y4 disoccupati	1,14	0,00	6,64	0,46	40,3
Y5 in cerca 1 occupazione	2,75	0,00	6,16	0,77	27,9
Y6 occupati agricoltura	8,03	0,00	78,05	11,28	140,4
Y7 occupati industria	48,64	3,85	85,96	11,22	23,1
Y8 occupati altre attività	43,33	7,32	91,38	11,63	26,8
Y9 imprenditori e liberi prof.	1,29	0,00	8,33	0,54	42,1
Y10 lavoratori dipendenti	19,91	3,85	36,86	3,78	19,0
		TOSCANA			
Y1 laurea	2,65	0,17	7,20	1,84	69,5
Y2 lic. scuola media inferiore	20,75	12,46	27,34	2,40	11,5
Y3 occupati	38,14	23,70	47,05	4,00	10,5
Y4 disoccupati	1,18	0,00	4,27	0,46	38,8
Y5 incerca 1 occupazione	2,67	1,30	5,09	0,54	20,4
Y6 occupati agricoltura	6,77	0,83	48,94	7,37	108,9
Y7 occupati industria	42,94	15,58	79,33	14,09	32,8
Y8 occupati altre attività	50,29	19,58	78,72	15,32	30,5
Y9 imprenditori e liberi prof.	1,65	0,00	3,85	0,72	43,7
Y10 lavoratori dipendenti	18,99	10,23	31,39	4,12	21,7
		CALABRIA			
Y1 laurea	2,31	0,12	5,41	1,49	64,5
Y2 lic. scuola media inferiore	17,48	10,47	23,97	2,71	15,5
Y3 occupati	25,35	12,28	43,03	3,64	14,4
Y4 disoccupati	3,58	0,34	24,67	3,21	89,6
Y5 in cerca 1 occupazione	6,26	2,16	14,89	1,26	20,2
Y6 occupati agricoltura	23,27	1,76	81,93	17,04	73,2
Y7 occupati industria	28,17	5,86	59,42	9,74	34,6
Y8 occupati altre attività	48,56	11,43	91,47	19,02	39,2
Y9 imprenditori e liberi prof.	0,68	0,00	1,88	0,35	51,1
Y10 lavoratori dipendenti	17,15	8,25	38,76	5,34	31,1

Prospetto 5 - Matrice di correlazione delle variabili «obiettivo»

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
					PIEMONTE					
Y1	1,00	0,55	-0,21	0,28	0,36	-0,47	-0,40	0,84	0,59	-0,33
Y2	0,55	1,00	0,07	0,45	0,69	-0,72	0,16	0,55	0,36	0,27
Y3	-0,21	0,07	1,00	-0,07	-0,07	0,07	0,27	-0,33	0,05	0,39
Y4	0,28	0,45	-0,07	1,00	0,34	-0,58	0,22	0,35	0,15	0,38
Y5	0,36	0,69	-0,07	0,34	1,00	-0,51	0,17	0,33	0,06	0,31
Y6	-0,47	-0,72	0,07	-0,58	-0,51	1,00	-0,47	-0,52	-0,36	-0,53
Y7	-0,40	0,16	0,27	0,22	0,17	-0,47	1,00	-0,51	-0,19	0,85
Y8	0,84	0,55	-0,33	0,35	0,33	-0,52	-0,51	1,00	0,53	-0,31
Y9	0,59	0,36	0,05	0,15	0,06	-0,36	-0,19	0,53	1,00	-0,23
Y10	-0,33	0,27	0,39	0,38	0,31	-0,53	0,85	-0,31	-0,23	1,00
					TOSCANA					
Y1	1,00	0,66	-0,18	-0,12	0,19	-0,41	-0,70	0,84	0,50	-0,76
Y2	0,66	1,00	-0,16	0,09	0,37	-0,59	-0,44	0,69	0,49	-0,53
Y3	-0,18	-0,16	1,00	-0,34	-0,61	-0,13	0,55	-0,44	0,34	0,61
Y4	-0,12	0,09	-0,34	1,00	0,24	0,01	-0,08	0,07	-0,10	-0,07
Y5	0,19	0,37	-0,61	0,24	1,00	-0,13	-0,36	0,39	-0,13	-0,39
Y6	-0,41	-0,59	-0,13	0,01	-0,13	1,00	-0,09	-0,40	-0,51	0,14
Y7	-0,70	-0,44	0,55	-0,08	-0,36	-0,09	1,00	-0,88	-0,15	0,83
Y8	0,84	0,69	-0,44	0,07	0,39	-0,40	-0,88	1,00	0,38	-0,83
Y9	0,50	0,49	0,34	-0,10	-0,13	-0,51	-0,15	0,38	1,00	-0,31
Y10	-0,76	-0,53	0,61	-0,07	-0,39	0,14	0,83	-0,83	-0,31	1,00
					CALABRIA					
Y1	1,00	0,63	0,38	-0,40	0,26	-0,71	-0,51	0,90	0,80	-0,68
Y2	0,63	1,00	0,19	-0,23	0,42	-0,53	-0,39	0,68	0,53	-0,45
Y3	0,38	0,19	1,00	-0,38	-0,14	-0,13	-0,21	0,22	0,43	0,05
Y4	-0,40	-0,23	-0,38	1,00	-0,23	0,62	-0,06	-0,53	-0,36	0,70
Y5	0,26	0,42	-0,14	-0,23	1,00	-0,26	-0,17	0,32	0,08	-0,37
Y6	-0,71	-0,53	-0,13	0,62	-0,26	1,00	-0,07	-0,86	-0,66	0,82
Y7	-0,51	-0,39	-0,21	-0,06	-0,17	-0,07	1,00	-0,45	-0,35	0,18
Y8	0,90	0,68	0,22	-0,53	0,32	-0,86	-0,45	1,00	0,77	-0,83
Y9	0,80	0,53	0,43	-0,36	0,08	-0,66	-0,35	0,77	1,00	-0,57
Y10	-0,68	-0,45	0,05	0,70	-0,37	0,82	0,18	-0,83	-0,57	1,00

11. ANALISI DEI RISULTATI DELLA SPERIMENTAZIONE

Prima di passare ad esaminare l'efficienza delle diverse stratificazioni prese in esame è opportuno soffermare l'attenzione sul criterio adottato per l'individuazione dei comuni autorappresentativi, che è quello attualmente utilizzato per il campione delle forze di lavoro.

Come si evince dai valori riportati nel prospetto 6, la percentuale dei comuni autorappresentativi sul totale dei comuni varia notevolmente da regione a regione, così come varia la percentuale di po-

polazione in essi residente e ciò comporta una diversificazione della struttura dei campioni che vengono generati. Infatti, per le tre regioni considerate si ha:

— per il Piemonte un campionamento casuale semplice per circa il 52% della popolazione residente nei 30 comuni autorappresentativi e un campionamento a due stadi per il restante 48% della popolazione residente nei 1179 comuni non autorappresentativi;

— per la Toscana un campionamento casuale semplice per circa il 61% della popolazione residente nei 35 comuni autorappresentativi e un campionamento a due stadi per il restante 39% della popolazione residente nei 252 comuni non autorappresentativi;

— per la Calabria un campionamento casuale semplice per circa il 33% della popolazione residente nei 12 comuni autorappresentativi e un campionamento a due stadi per il restante 67% della popolazione residente nei 397 comuni non autorappresentativi.

Risulta evidente che se per la determinazione dei comuni autorappresentativi si adotta una soglia di popolazione costante o una soglia variabile che tenga conto esclusivamente di considerazioni pratiche (come il numero minimo d'interviste d'assegnare a ciascun rilevatore), il disegno campionario può risultare più efficiente per alcuni domini territoriali e meno per altri. È quindi necessario che nella procedura generalizzata sia introdotta la possibilità di calcolare il valore «ottimo» della soglia all'interno di ciascun dominio territoriale, in funzione dei livelli di precisione delle stime, dei costi in primo e in secondo stadio e della distribuzione dei comuni per ampiezza demografica.

Prospetto 6 - Numero di comuni e popolazione per tipo di comune

TIPO DI COMUNE	CIFRE ASSOLUTE		% SUL TOTALE	
	Comuni	Popolazione	Comuni	Popolazione
Piemonte				
Autorappresentativi	30	2.318.232	2,48	51,76
Non autorappresentativi	1179	2.160.799	97,52	48,24
Totale	**1209**	**4.479.031**	**100,00**	**100,00**
Toscana				
Autorappresentativi	35	2.176.500	12,20	60,78
Non autorappresentativi	252	1.404.551	87,80	39,72
Totale	**287**	**3.581.051**	**100,00**	**100,00**
Calabria				
Autorappresentativi	12	688.265	2,93	33,39
Non autorappresentativi	397	1.372.917	97,07	66,61
Totale	**409**	**2.061.182**	**100,00**	**100,00**

11.1 Efficienza dei metodi di stratificazione sperimentati

Indicando con S1 e S2 due generiche stratificazioni a confronto e con V(S1) e V(S2) le corrispondenti varianze campionarie dello stimatore di un totale calcolate tramite la (2), l'efficienza relativa della stratificazione S1 rispetto alla S2 è stata misurata mediante il rapporto V(S2)/V(S1). Pertanto la stratificazione S1 risulterà più vantaggiosa di S2 se il rapporto assumerà valori maggiori di uno e meno vantaggiose se assumerà valori inferiori all'unità.

Elaborando i files con le varianze campionarie relative a ciascuna delle stratificazioni sperimentate, sono stati effettuati due diversi tipi di confronti:

a) confronti fra stratificazioni che utilizzano le stesse variabili ma metodi diversi per il raggruppamento dei comuni;

b) confronti fra stratificazioni effettuate con variabili diverse.

I risultati dei confronti sono riportati nelle tavole 1-7 poste alla fine del paragrafo.

1. *Confronti fra le stratificazioni che utilizzano le variabili relative all'attività economica*

Nella tavola 1 sono riportati i valori dei rapporti tra le varianze relative alle stratificazioni basate sulla cluster analysis delle percentuali di popolazione in condizione professionale per ramo di attività

economica effettuate a livello di provincia (AE2), di zona altimetrica all'interno della provincia (AE3) e di settore (AE4) e la varianza della stratificazione attualmente utilizzata per il campione delle forze di lavoro.

Analizzando i risultati relativi al campione base si ha che per le variabili Y6, Y7 e Y8, che sono quelle utilizzate per la stratificazione, i procedimenti basati sulla cluster analysis risultano nettamente più efficienti di quello attualmente utilizzato che è basato sull'attività economica prevalente. Anche per le altre variabili si riscontra, quasi sempre, una maggiore efficienza, anche se più attenuata, delle stratificazioni basate sulla cluster analysis.

La suddivisione dei comuni per zona altimetrica o per settore statistico prima della cluster, non sembra comportare variazioni apprezzabili nelle varianze campionarie.

Occorre evidenziare che non è stato possibile ripetere i confronti per il campione ridotto in quanto la stratificazione AE1 porta ad un numero fisso di strati, uguale a quello del campione base. Una riduzione del numero degli strati mediante un loro accorpamento, ottenuto facendo saltare uno o più caratteri di stratificazione, modifica la stratificazione stessa e rende i risultati dei confronti difficilmente interpretabili. Inoltre, sempre per il campione ridotto, non si è potuto procedere alla stratificazione AE4 perché alcune volte il numero dei settori è risultato maggiore del numero degli strati che dovevano essere formati.

2. *Confronti fra le stratificazioni che utilizzano la popolazione residente*

Nelle tavole 2 e 3 sono riportati i valori dell'efficienza della stratificazione POP2 (classi di popolazione residente con uguale ammontare della popolazione) rispetto alle altre quattro stratificazioni basate sulla popolazione residente, rispettivamente per il campione base e per quello ridotto.

La stratificazione POP2 risulta più efficiente delle altre stratificazioni per tutte le regioni e le variabili considerate (ad eccezione della POP5 per la Calabria). In particolare i confronti mostrano che l'efficienza è risultata più marcata rispetto alla stratificazione POP1 basata su classi di popolazione con uguale numero di comuni, e più attenuata nei confronti delle stratificazioni basate sulla cluster analysis della popolazione residente (POP3, POP4 e POP5).

Se si confrontano tra di loro i valori relativi a queste tre ultime stratificazioni si evidenzia che per il Piemonte e la Toscana le tre stratificazioni non portano a variazioni apprezzabili nelle varianze campionarie, mentre per la Calabria la cluster analysis effettuata all'interno dei settori (POP5) risulta più efficiente di quelle effettuate all'interno delle province e delle zone altimetriche.

La maggiore efficienza della stratificazione POP2 rispetto alle altre stratificazioni basate sulla popolazione residente è confermata anche dall'esame dei risultati relativi al campione ridotto. Come già detto in precedenza la stratificazione basata sulla cluster analysis all'interno dei settori (POP5) non può essere effettuata per il campione ridotto, in quanto in alcuni casi il numero dei settori è più elevato del numero degli strati da formare.

3. *Confronti fra le stratificazioni che utilizzano le prime due componenti principali*

Nelle tavole 4 e 5 sono riportati i valori dell'efficienza della stratificazione PR1 (cluster analysis delle prime due componenti principali effettuata all'interno di ciascuna provincia) rispetto a PR2 e PR3 per le quali la cluster analysis è stata effettuata all'interno delle zone altimetriche e ai settori statistici di ciascuna provincia. I risultati non sono omogenei per le tre regioni, infatti mentre per il Piemonte e la Calabria la suddivisione dei comuni per zona altimetrica e per settore prima della cluster analysis porta ad una stratificazione più efficiente, per la Toscana risulta più efficiente la PR1.

Il confronto condotto sul campione ridotto riguarda solo la PR2 e conferma i risultati trovati nel campione base.

4. *Confronti fra stratificazioni effettuate con variabili diverse*

Un'ultima serie di confronti (tavola 6) è stata effettuata per valutare l'efficienza della stratificazione basata sull'uguale ammontare della popolazione negli strati (POP2), che si è dimostrata la più vantaggiosa tra quelle basate sulla popolazione residente, rispetto alle due stratificazioni che utilizzano la cluster analysis a livello provinciale delle percentuali di addetti per settore di attività economica (AE2) e delle prime due componenti principali (PR1).

La stratificazione AE2 risulta notevolmente più efficiente per le variabili Y6, Y7, Y8 che sono quelle utilizzate da AE2 per la formazione degli strati, mentre in genere è leggermente meno efficiente per le altre variabili.

Per il Piemonte la stratificazione POP2 risulta più efficiente di PR1 per tutte le variabili, tranne che per Y1 e Y6 per le quali le due stratificazioni si equivalgono. Per la Toscana POP2 è più efficiente per una sola variabile e meno efficiente per altre quattro, mentre per cinque variabili non si riscontrano differenze apprezzabili. Infine per la Calabria POP2 risulta più efficiente per cinque variabili e meno efficiente per tre.

Un'analisi condotta per variabili evidenzia che POP2 risulta sempre più efficiente di PR1 per le variabili Y2 (licenza di scuola media inferiore), Y4 (disoccupati) e Y5 (in cerca di prima occupazione), mentre risulta sempre meno efficiente per la Y6 (occupati in agricoltura) e la Y1 (laureati).

11.2 Praticabilità dei metodi di stratificazione sperimentati

La scelta di un metodo di stratificazione non può essere basata esclusivamente su considerazioni relative all'efficienza, ma deve tenere conto anche della sua praticabilità, e questa spesso è legata al numero dei comuni e alla popolazione residente negli strati formati.

Tutti i metodi di stratificazione considerati, ad esclusione di quelli basati sulla popolazione residente, presentano l'inconveniente di condurre a strati costituiti molte volte da uno o pochi comuni anche di piccole dimensioni, o a strati molto numerosi e con una notevole variabilità dell'ampiezza demografica.

Per assicurare un numero minimo di interviste per rilevatore, anche utilizzando un solo rilevatore per strato, occorre che nello strato stesso sia compresa una popolazione $P > k \cdot m/f$, dove k è il numero minimo di interviste, m il numero medio di componenti per famiglia ed f è il tasso di campionamento finale. Così per $f = 1/200$ e $m = 3$, volendo assegnare $k = 15$ famiglie da intervistare per rilevatore si ha: $P > 15 \cdot 15 \cdot 3 \cdot 200 = 9000$.

Un secondo vincolo riguarda l'ampiezza demografica minima dei comuni all'interno di uno strato, che non può essere inferiore al numero delle persone che devono essere intervistate nello strato: $\min A > f \cdot P$, dove P è la popolazione dello strato. Così uno strato con 100.000 abitanti ed un tasso di campionamento $f = 1/200$ non può comprendere comuni con una popolazione inferiore a 500 abitanti.

Qualche volta questo inconveniente può presentarsi anche per la stratificazione effettuata in base alla popolazione residente con uguale ammontare negli strati (POP2). Quando ciò si verifica, in ciascuna area di stratificazione ne risulta interessato soltanto il primo strato, quello che contiene il comune con popolazione minima.

Nella tavola 7 è riportata la popolazione residente, il numero di persone che devono essere intervistate e la popolazione minima nel primo strato di ciascuna provincia della stratificazione POP2. Dall'esame dei dati si evidenzia che soltanto per due strati del Piemonte (11001 e 21001) si riscontra la presenza di comuni con popolazione inferiore a quella che deve essere campionata.

L'inconveniente può essere superato spezzando gli strati in due o più substrati in modo che la popolazione residente in ciascuno di essi sia inferiore al rapporto fra la popolazione minima e il tasso di campionamento.

Questo modo di procedere fa aumentare di qualche unità il numero degli strati, e quindi il numero dei comuni da campionare, e per i substrati così formati può accadere che agli intervistatori venga assegnato un numero d'interviste inferiore a quello minimo stabilito.

Un'altra strada che può essere seguita, ma che è completamente da esplorare, è quella della costruzione di unità di primo stadio artificiali, mediante aggregazione di due o più comuni contigui, in modo da garantire un'ampiezza demografica minima ad ogni unità di primo stadio.

11.3 Considerazioni finali

La procedura generalizzata predisposta permette all'utilizzatore di ottenere rapidamente una stratificazione dei comuni italiani, consentendo una notevole flessibilità nelle varie fasi del processo di stratificazione e nello stesso tempo fornendo in output una serie di informazioni che lo possono indirizzare nella scelta del procedimento migliore sia da un punto di vista dell'efficienza che della praticabilità.

I programmi SAS predisposti hanno consentito, con poche istruzioni, di poter effettuare analisi anche piuttosto complesse (come quella in componenti principali e la cluster analysis), e si sono di-

mostrati molto efficaci anche in tutte le altre fasi previste dalla procedura. Tuttavia, per un'utilizzazione generalizzata, è necessario che sia messo a punto di un programma interattivo, che, mediante pannelli, guidi l'utente nelle diverse fasi del procedimento.

Da quanto emerso dai risultati della sperimentazione è apparso evidente che la stratificazione basata sulla popolazione residente presenta notevoli punti a favore per quanto riguarda gli aspetti legati alla praticabilità, ed inoltre, non risulta meno efficiente delle stratificazioni basate su altre variabili, a meno che queste ultime non coincidano con le variabili oggetto d'indagine (o non siano con queste fortemente correlate).

Occorre anche tener presente che la popolazione residente presenta il vantaggio di poter essere aggiornata annualmente a differenza delle altre variabili i cui valori, a livello comunale, si hanno soltanto in corrispondenza dei censimenti, e quindi spesso risultano obsoleti.

In definitiva la procedura generalizzata può prevedere per default la stratificazione dei comuni mediante classi di popolazione con uguale ammontare della popolazione, e l'impiego di uno degli altri metodi mediante opportuna specificazione da parte dell'utilizzatore.

Tavola 1 - Efficienza della stratificazione «AE1» rispetto alle altre stratificazioni basate sull'attività economica *(campione base, tasso di campionamento f = 1/200)*

VARIABILI OBIETTIVO	AE2	AE3	AE4
	PIEMONTE		
Y1 laurea	0,9981	1,1057	0,9803
Y2 lic. scuola media inferiore	0,9777	1,0410	0,9281
Y3 occupati	0,9940	1,1594	0,9720
Y4 disoccupati	0,9689	1,0453	0,9709
Y5 in cerca 1^ occupazione	0,9803	1,0331	0,9862
Y6 occupati agricoltura	0,4289	0,3694	0,5740
Y7 occupati industria	0,7337	0,8424	0,7543
Y8 occupati altre attività	0,8513	0,8377	0,8993
Y9 imprenditori e liberi prof.	0,9764	1,0502	0,9941
Y10 lavoratori dipendenti	0,8298	0,9801	0,8643
	TOSCANA		
Y1 laurea	0,9995	0,9864	0,9897
Y2 lic. scuola media inferiore	0,9957	0,9821	0,9844
Y3 occupati	0,9794	1,0336	0,9767
Y4 disoccupati	0,9731	0,9736	0,9584
Y5 in cerca 1^ occupazione	1,0032	1,0006	0,9949
Y6 occupati agricoltura	0,6841	0,5744	0,7252
Y7 occupati industria	0,7251	0,7093	0,7450
Y8 occupati altre attività	0,8625	0,8349	0,8780
Y9 imprenditori e liberi prof.	0,9886	0,9954	0,9885
Y10 lavoratori dipendenti	0,9716	0,9377	0,9542
	CALABRIA		
Y1 laurea	0,9053	0,8829	0,8887
Y2 lic. scuola media inferiore	1,0112	1,0404	0,9867
Y3 occupati	1,0390	1,0569	0,9718
Y4 disoccupati	0,7985	0,8709	0,8479
Y5 in cerca 1^ occupazione	1,0175	1,0132	0,9853
Y6 occupati agricoltura	0,6776	0,5926	0,7032
Y7 occupati industria	0,9007	0,8412	0,9488
Y8 occupati altre attività	0,7776	0,7418	0,7631
Y9 imprenditori e liberi prof.	0,9816	0,9781	0,9728
Y10 lavoratori dipendenti	0,9433	0,9118	0,8433

Tavola 2 - Efficienza della stratificazione «POP2» rispetto alle altre stratificazioni basate sulla popolazione *(campione base, tasso di campionamento f = 1/200)*

VARIABILI OBIETTIVO	POP1	POP3	POP4	POP5
	PIEMONTE			
Y1 laurea	1,2071	1,0025	1,0103	1,0828
Y2 lic. scuola media inferiore	1,1345	1,0315	1,0795	1,0849
Y3 occupati	1,0909	1,0472	1,0267	1,0317
Y4 disoccupati	1,0926	1,0187	1,0264	1,0495
Y5 in cerca 1^ occupazione	1,0576	1,0081	1,0172	1,0246
Y6 occupati agricoltura	1,3075	1,2416	1,3047	1,1333
Y7 occupati industria	1,4318	1,1041	1,0828	1,0165
Y8 occupati altre attività	1,3309	1,0262	1,0074	1,0728
Y9 imprenditori e liberi prof.	1,0848	1,0154	1,0250	1,0252
Y10 lavoratori dipendenti	1,3059	1,0530	1,0910	1,0744
	TOSCANA			
Y1 laurea	1,0531	1,0225	1,0263	1,0307
Y2 lic. scuola media inferiore	1,0362	1,0187	1,0288	1,0250
Y3 occupati	1,0448	1,0280	1,0082	1,0228
Y4 disoccupati	1,0489	1,0054	1,0042	1,0220
Y5 in cerca 1^ occupazione	1,0068	1,0037	1,0003	1,0054
Y6 occupati agricoltura	1,1799	1,0359	1,0402	1,0993
Y7 occupati industria	1,2652	1,1159	1,1914	1,1276
Y8 occupati altre attività	1,1389	1,0763	1,1010	1,1033
Y9 imprenditori e liberi prof.	1,0526	1,0139	1,0411	1,0237
Y10 lavoratori dipendenti	1,0921	1,0503	1,0358	1,0425
	CALABRIA			
Y1 laurea	1,2573	1,0016	1,0104	1,0034
Y2 lic. scuola media inferiore	1,1236	1,0264	1,0137	1,0234
Y3 occupati	1,1212	1,0416	1,0562	0,9557
Y4 disoccupati	1,4165	1,2304	0,9568	1,0688
Y5 in cerca 1^ occupazione	1,0706	1,0348	1,0435	1,0055
Y6 occupati agricoltura	1,5924	1,2442	0,9896	0,8073
Y7 occupati industria	1,2889	1,0453	1,0334	0,9727
Y8 occupati altre attività	1,6460	1,0414	1,0413	0,9801
Y9 imprenditori e liberi prof.	1,0485	1,0015	1,0018	0,9963
Y10 lavoratori dipendenti	1,4079	1,1580	1,0086	0,9062

Tavola 3 - Efficienza della stratificazione «POP2» rispetto alle altre stratificazioni basate sulla popolazione *(campione base, tasso di campionamento f = 1/500)*

VARIABILI OBIETTIVO	POP1	POP3	POP4	POP5
	PIEMONTE			
Y1 laurea	1,1475	1,0404	1,0259	—
Y2 lic. scuola media inferiore	1,1295	1,0373	1,0555	—
Y3 occupati	1,0795	1,0509	1,0289	—
Y4 disoccupati	1,0819	1,0210	1,0231	—
Y5 in cerca 1^ occupazione	1,0410	1,0085	1,0170	—
Y6 occupati agricoltura	1,4492	1,4116	1,3399	—
Y7 occupati industria	1,3622	1,1284	1,0728	—
Y8 occupati altre attività	1,2448	1,0741	1,0241	—
Y9 imprenditori e liberi prof.	1,0551	1,0298	1,0234	—
Y10 lavoratori dipendenti	1,2569	1,0957	1,0808	
	TOSCANA			
Y1 laurea	1,0565	1,0355	1,1027	—
Y2 lic. scuola media inferiore	1,0369	1,0143	1,0504	—
Y3 occupati	1,0386	1,0226	1,0103	—
Y4 disoccupati	1,0422	1,0199	1,0448	—
Y5 in cerca 1^ occupazione	1,0052	1,0006	1,0124	—
Y6 occupati agricoltura	1,0814	0,9962	1,3179	—
Y7 occupati industria	1,2243	1,1244	1,3492	—
Y8 occupati altre attività	1,1671	1,0911	1,2491	—
Y9 imprenditori e liberi prof.	1,0428	1,0203	1,0269	—
Y10 lavoratori dipendenti	1,0775	1,0311	1,0875	—
	CALABRIA			
Y1 laurea	1,2204	1,0289	0,9979	—
Y2 lic. scuola media inferiore	1,1268	1,0649	1,0272	—
Y3 occupati	1,1195	1,1279	1,0965	—
Y4 disoccupati	1,3311	1,3464	1,0796	—
Y5 in cerca 1^ occupazione	1,0630	1,0644	1,0510	—
Y6 occupati agricoltura	1,4888	1,3794	1,0793	—
Y7 occupati industria	1,2987	1,1485	1,1451	—
Y8 occupati altre attività	1,5143	1,0882	1,0219	—
Y9 imprenditori e liberi prof.	1,0606	1,0191	1,0213	—
Y10 lavoratori dipendenti	1,3297	1,2288	1,0721	—

Tavola 4 - Efficienza della stratificazione «PR1» rispetto alle altre stratificazioni basate sulle componenti principali *(campione base, tasso di campionamento f = 1/200)*

VARIABILI OBIETTIVO	PR2	PR3
PIEMONTE		
Y1 laurea	0,9564	0,9971
Y2 lic. scuola media inferiore	0,9403	0,9733
Y3 occupati	0,8791	0,8835
Y4 disoccupati	0,9247	0,9428
Y5 in cerca 1^ occupazione	0,9617	0,9736
Y6 occupati agricoltura	0,7864	0,9209
Y7 occupati industria	0,7821	0,8438
Y8 occupati altre attività	0,8466	0,9752
Y9 imprenditori e liberi prof.	0,9696	0,9933
Y10 lavoratori dipendenti	0,9192	1,0108
TOSCANA		
Y1 laurea	1,0138	1,0014
Y2 lic. scuola media inferiore	1,0183	1,0011
Y3 occupati	0,9866	1,0171
Y4 disoccupati	1,0048	0,9664
Y5 in cerca 1^ occupazione	1,0007	1,0013
Y6 occupati agricoltura	1,1116	1,1101
Y7 occupati industria	1,0823	1,0364
Y8 occupati altre attività	1,0271	1,0182
Y9 imprenditori e liberi prof.	1,0231	0,9988
Y10 lavoratori dipendenti	1,0914	1,0258
CALABRIA		
Y1 laurea	0,9991	1,0096
Y2 lic. scuola media inferiore	0,9538	0,9504
Y3 occupati	0,9473	0,8999
Y4 disoccupati	1,0041	0,9727
Y5 in cerca 1^ occupazione	0,9723	0,9468
Y6 occupati agricoltura	0,9164	0,9361
Y7 occupati industria	0,9409	0,8653
Y8 occupati altre attività	1,0022	1,0898
Y9 imprenditori e liberi prof.	0,9978	0,9973
Y10 lavoratori dipendenti	0,9491	0,9169

Tavola 5 - Efficienza della stratificazione «PR1» rispetto alle altre stratificazioni basate sulle componenti principali *(campione base, tasso di campionamento f = 1/500)*

VARIABILI OBIETTIVO	PR2	PR3
PIEMONTE		
Y1 laurea	0,9978	—
Y2 lic. scuola media inferiore	0,9629	—
Y3 occupati	0,9534	—
Y4 disoccupati	0,9579	—
Y5 in cerca 1^ occupazione	0,9815	—
Y6 occupati agricoltura	0,8653	—
Y7 occupati industria	0,9028	—
Y8 occupati altre attività	0,9504	—
Y9 imprenditori e liberi prof.	0,9893	—
Y10 lavoratori dipendenti	0,9802	—
TOSCANA		
Y1 laurea	1,1134	—
Y2 lic. scuola media inferiore	1,0389	—
Y3 occupati	1,0402	—
Y4 disoccupati	1,0140	—
Y5 in cerca 1^ occupazione	1,0101	—
Y6 occupati agricoltura	1,5118	—
Y7 occupati industria	1,6173	—
Y8 occupati altre attività	1,3024	—
Y9 imprenditori e liberi prof.	1,0114	—
Y10 lavoratori dipendenti	1,1238	—
CALABRIA		
Y1 laurea	1,0000	—
Y2 lic. scuola media inferiore	0,8944	—
Y3 occupati	0,9055	—
Y4 disoccupati	0,8692	—
Y5 in cerca 1^ occupazione	0,8553	—
Y6 occupati agricoltura	0,9485	—
Y7 occupati industria	0,8642	—
Y8 occupati altre attività	0,8948	—
Y9 imprenditori e liberi prof.	0,9877	—
Y10 lavoratori dipendenti	0,9824	—

Tavola 6 - Efficienza della stratificazione «POP2» rispetto alle stratificazioni AE2 E PR1
(campione base, tasso di campionamento f = 1/200)

VARIABILI OBIETTIVO	AE2	PR1
PIEMONTE		
Y1 laurea	0,9934	0,9996
Y2 lic. scuola media inferiore	1,1017	1,0694
Y3 occupati	1,0302	1,0706
Y4 disoccupati	1,0384	1,0674
Y5 in cerca 1^ occupazione	1,0176	1,0277
Y6 occupati agricoltura	0,6851	0,9964
Y7 occupati industria	0,9052	1,1166
Y8 occupati altre attività	0,9715	1,0844
Y9 imprenditori e liberi prof.	1,0078	1,0153
Y10 lavoratori dipendenti	0,9820	1,0138
TOSCANA		
Y1 laurea	1,0002	0,9982
Y2 lic. scuola media inferiore	1,0155	1,0090
Y3 occupati	0,9993	0,9917
Y4 disoccupati	1,0027	1,0148
Y5 in cerca 1^ occupazione	1,0035	1,0021
Y6 occupati agricoltura	0,9089	0,9480
Y7 occupati industria	0,8688	0,8934
Y8 occupati altre attività	0,9600	0,9748
Y9 imprenditori e liberi prof.	1,0128	1,0021
Y10 lavoratori dipendenti	1,0022	0,9764
CALABRIA		
Y1 laurea	1,0160	0,9977
Y2 lic. scuola media inferiore	1,0320	1,0314
Y3 occupati	1,0500	1,0660
Y4 disoccupati	0,9650	1,0220
Y5 in cerca 1^ occupazione	1,0262	1,0275
Y6 occupati agricoltura	0,6260	0,8294
Y7 occupati industria	0,9434	1,0717
Y8 occupati altre attività	0,9411	0,9257
Y9 imprenditori e liberi prof.	1,0036	0,9977
Y10 lavoratori dipendenti	0,9221	0,9319

Tavola 7 - Popolazione residente, numero di persone da intervistare e popolazione minima nel primo strato di ciascuna provincia

PROVINCE	Codice di strato	Popolazione residente strato	Persone da intervistare	Popolazione residente minima
	PIEMONTE			
1. Torino	11001	27.765	139	32
2. Vercelli	21001	29.101	146	78
3. Novara	31001	32.228	161	145
4. Cuneo	41001	14.925	75	77
5. Asti	51001	19.842	99	144
6. Alessandria	61001	13.772	69	129
	TOSCANA			
45. Massa	451001	13.284	66	966
46. Lucca	461001	28.620	143	562
47. Pistoia	471001	34.741	174	821
48. Firenze	481001	23.482	117	1.089
49. Livorno	491001	23.478	119	395
50. Pisa	501001	24.513	123	588
51. Arezzo	511001	44.142	221	514
52. Siena	521001	14.367	72	1.090
53. Grosseto	531001	14.689	73	1.239
	CALABRIA			
78. Cosenza	781001	17.440	81	461
79. Catanzaro	791001	16.012	80	627
80. Reggio di Calabria	801001	19.269	96	563

APPENDICE

ELENCO DELLE VARIABILI ESTRATTE DALL'ARCHIVIO DEI DATI DI CENSIMENTO

Variabili territoriali e codici identificativi

REG = codice di regione
PROV = codice di provincia
COM = codice di comune
NOME = nome del comune
USL = codice di USL
ALT = codice di zona altimetrica
SETTORE = codice di settore statistico

Popolazione residente e superficie (tav. 1)

Y1 = Superficie
Y2 = Popolazione maschile
Y3 = Popolazione femminile

Popolazione residente per stato civile (tav. 3)

M	F	
Y4	Y9	= celibi o nubili
Y5	Y10	= coniugati
Y6	Y11	= separati
Y7	Y12	= vedovi
Y8	Y12	= divorziati

Popolazione residente per classi quinquennali di età (tav. 4)

Y14 = Popolazione da 0 a 4 anni
Y15 = Popolazione da 5 a 9 anni
Y16 = Popolazione da 10 a 14 anni
Y17 = Popolazione da 15 a 19 anni
Y18 = Popolazione da 20 a 24 anni
Y19 = Popolazione da 25 a 29 anni
Y20 = Popolazione da 30 a 34 anni
Y21 = Popolazione da 35 a 39 anni
Y22 = Popolazione da 40 a 44 anni
Y23 = Popolazione da 45 a 49 anni
Y24 = Popolazione da 50 a 54 anni
Y25 = Popolazione da 55 a 59 anni
Y26 = Popolazione da 60 a 64 anni

Y27 = Popolazione da 65 a 69 anni
Y28 = Popolazione da 70 a 74 anni
Y29 = Popolazione da 75 ed oltre

Popolazione residente per particolari classi di età (tav. 4)

Y30 = Popolazione da 0 a 2 anni
Y31 = Popolazione da 3 a 5 anni
Y32 = Popolazione da 6 a 10 anni
Y33 = Popolazione da 11 a 13 anni
Y34 = Popolazione da 14 a 17 anni
Y35 = Popolazione di 18 anni
Y36 = Popolazione da 19 a 20 anni
Y37 = Popolazione da 21 a 24 anni

Popolazione residente da 6 anni in poi per titolo di studio (tav. 5)

Y38 = Laurea
Y39 = Diploma
Y40 = Licenza media inferiore
Y41 = Licenza elementare
Y42 = Alfabeti privi di titolo di studio
Y43 = Analfabeti

Popolazione residente che frequenta corsi regolari di studio o corsi di formazione professionale per classi di età (tav. 6)

Y44 = Corsi regolari classe di età 6-13
Y45 = Corsi regolari classe di età 14-18
Y46 = Corsi regolari classe di età 19 e più
Y47 = Corsi di formazione professionale classe di età 14-16
Y48 = Corsi di formazione professionale classe di età 17-19
Y49 = Corsi di formazione professionale classe di età 20 e più

Popolazione residente attiva per condizione (tav. 7)

Y50 = Occupati
Y51 = Disoccupati
Y52 = In cerca di prima occupazione

Popolazione residente attiva in condizione professionale per sesso e ramo di attività economica (tav. 8)

Y53 = Ramo 0
Y54 = Ramo 1
Y55 = Ramo 2

Y56 = Ramo 3
Y57 = Ramo 4
Y58 = Ramo 5
Y59 = Ramo 6
Y60 = Ramo 7
Y61 = Ramo 8
Y62 = Ramo 9

Popolazione residente attiva in condizione professionale per posizione nella professione (tav. 9)

Y63 = Imprenditori e liberi professionisti
Y64 = Lavoratori in proprio

Y71 = casalinghe > 14 anni
Y72 = studenti > 14 anni
Y73 = ritirati dal lavoro > 14 anni
Y74 = altri > 14 anni

Popolazione residente che rientra giornalmente nella propria dimora abituale, secondo il luogo di lavoro o di studio (Tav. 12)

Y75 = occupati nello stesso comune
Y76 = fuori del comune
Y77 = scolari e studenti nello stesso comune
Y78 = fuori del comune
Y79 = Persone che frequentano un corso di formazione professionali nello stesso comune
Y80 = fuori del comune

Famiglie residenti per ampiezza della famiglia (Tav. 13)

Y81 = 1 componente
Y82 = 2 componenti
Y83 = 3 componenti
Y84 = 4 componenti
Y85 = 5 componenti
Y86 = 6 componenti
Y87 = 7 componenti
Y88 = 8 e più

Famiglie residenti secondo la tipologia (Tav. 14)

Y89 = Coniugi
Y90 = di cui con altre persone
Y91 = Coniugi e figli

Y92 = di cui con altre persone
Y93 = un genitore e figli
Y94 = di cui con altre persone
Y95 = altro tipo di famiglia
Y96 = famiglie (totale)
Y97 = 8 componenti

Abitazioni occupate e non occupate e altri tipi di alloggio (Tav. 15)

Occupate
Y98 = N. abitazioni
Y99 = N. stanze
Y100 = Famiglie
Y101 = componenti

Non occupate
Y102 = N. abitazioni
Y103 = N. stanze

Altri tipi di alloggio
Y104 = Numero
Y105 = N. famiglie
Y106 = N. componenti

Abitazioni occupate per titolo di godimento (Tav. 16)

Proprietà
Y107 = n. abitazioni
Y108 = stanze
Y109 = famiglie
Y110 = componenti

Affitto
Y111 = n. abitazioni
Y112 = stanze
Y113 = famiglie
Y114 = componenti

Altro titolo
Y115 = n. abitazioni
Y116 = stanze
Y117 = famiglie
Y118 = componenti

Abitazioni occupate per servizio installato (Tav. 19)

Acqua potabile
Y119 = di acquedotto entro l'abitazione
Y120 = di acquedotto fuori l'abitazione

Y121 = pozzo o cisterna

Gabinetto

Y122 = nell'abitazione
Y123 = fuori dell'abitazione
Y124 = bagno
Y125 = elettricità

Riscaldamento

Y126 = Impianto fisso
Y127 = apparecchi singoli
Y128 = sforniti di acqua potabile e gabinetto

N.B. Le tavole indicate tra parentesi sono quelle dei fascicoli provinciali che contengono le variabili in oggetto.

RIFERIMENTI BIBLIOGRAFICI

BIGGERI L., CHIANDOTTO B. e GHILARDI G. (1977), «*Materiale di discussione dei primi risultati di una stratificazione dei comuni della Toscana*», Istat, Commissione per gli studi statistici ed econometrici interessanti la programmazione economica, doc. n. 58.

COCHRAN, W.G. (1961), «*Comparisson of Methods for Determining Stratum Boundaries*» Bulletin of the International Statistical Institute, 38 (2) Tokio, 345-358.

COCHRAN, W.G. (1977), «*Sampling Tecniques*», Wiley and Sons, New York.

DALENIUS, T. (1952), «*The Problem of Optimum Stratification in Special Type of Design*», Skandinavisk Aktuariedskrift, 35, 61-70.

DALENIUS, T. and GOURNEY, H. (1951), «*The problem of Optimum Stratification*», Skandinavisk Actuariedskrift, 34, 133-148.

DALENIUS, I. and HODGES, J.L. (1959), «*Minimum Variance Stratification*», Journal American Statistical Association, 54, 88-101.

EKMAN, G. (1959), «*An Approximation Useful in Univariate Stratification*», The Annals of Mathematical Statistics, 30, 219-229.

FABBRIS L. (1988), «*Campioni di numerosità due e tre per strato selezionati con probabilità variabili: valutazione empirica di alcune proprietà di stime di frequenza assolute*», Istat, Commissione di studio per la progettazione e l'applicazione dei campioni.

FABBRIS L. e ZANNELLA F. (1988), «*Schema della procedura generalizzata per la stratificazione e la selezione dei comuni delle indagini campionarie sulla popolazione*», Istat, Commissione di studio per la progettazione e l'applicazione dei campioni.

GAGGIOTTI M. e ZUCCHEGNA A. (1985), «*Procedura per la gestione dell'archivio dei dati comunali di censimento*», Istat, Commissione di studio per la progettazione e l'applicazione dei campioni.

GHOSH, S.P. (1963), «*Optimum Stratification Whith Two Characters*». Ann. Math. Statist., 34, 866-872.

GLASSER, G.J. (1962), «*On the Complete Coverage of Large Units In A Statistical Study*» Review of the International Statistical Institute, vol. 30, 28-32.

GOLDER P.A. and YEOMANS K.A. (1973), «*Use of cluster analysis for Stratification*», Appl. Statist. 22, pp. 213-219.

HANSEN, M.M. HURWITZ, W.G. and MADOW, U.G. (1953), «*Sample Survey Methods and Theory*», vol. I, John Wiley and Sons.

HESS, I., SETHI, V.K. and BALAKRISHANAN, T.R. (1966), «*Stratification: A Pratical Investigation*», Journal of American Statistical Association, 74-90.

HIDIROGLOU, M.A. (1979), «*On The Inclusion of Large Units in Simple Random Sampling*», American Statistical Association, Proceedings of the Section Research and Methods, 305-308.

HIDIROGLOU, M.A. (1986), «*The Construction of a Self-Representing Stratum of Large Units in Survey Design*», The American Statistician, vol. 40 n. 1, 27-31.

LAVALEE, P. and HIDIROGLOU, M.A. (1988), «*On the Stratification of Skewed Populations*», Survey Methodology, vol. 14 n. 1, 33-43.

ISTAT (1958), «*Rilevazioni campionarie delle forze di lavoro*», Metodi e Norme, Serie A, n. 3.

ISTAT (1958a), «*Circoscrizioni statistiche*», Metodi e Norme, Serie C, n. 1.

ISTAT (1969), «*Rilevazioni campionarie delle forze di lavoro*», Metodi e Norme, Serie A, n. 10.

ISTAT (1978), «*Rilevazioni campionarie delle forze di lavoro*», Metodi e Norme, Serie A, n. 15.

ISTAT (1986), «*Aggiornamento dell'elenco dei comuni al 31 dicembre 1985*», documento per uso interno.

JARQUE, C.M. (1981), «*A Solution of the Problem of Optimum Stratification in Multivariate Sampling*», Applied Statistics, 30, n. 2, 163-169.

KISH, L. (1965), «*Survey Sampling*», Wiley and Sons, New York.

MAHALANOBIS, P.C. (1952), «*Some Aspects of the Design of Sample Surveys*», Sankhia, 12, 1-7.

MURTHY, H.N. (1967), «*Sampling Theory and Methods*» Statistical Publishing Society, Calcutta.

NAPOLITANO P., RUSSO A. e ZANNELLA F. (1983), «*Calcolo, presentazione ed analisi degli errori di campionamento nell'indagine Istat sulle condizioni di salute della popolazione e sul ricorso ai servizi sanitari, novembre 1980*», SIS, Atti del Convegno, 1983 Trieste, 605-629.

NORLAND, R.E. (1983), «*An Efficient Algorithm Determining Strata Boundaries for Discrete Populations Using Ekman's Method*» American Statistical Association, Proceedings of the Computing Section, 174-176.

O'MUIRCHEARTHAIGH, C.A. (1977), «*Proximum Designs for Crude Sampling Frames*», Bull. Int. statist. Inst. 46, n. 3, 82-100.

RUSSO A. (1984), «*Piano della rilevazione campionaria ed errori di campionamento per l'indagine sulle vacanze e gli sports degli italiani nel 1982*», Istat Supplemento al Bollettino Mensile di Statistica n. 15, 8-12.

RUSSO A. (1985), «*Disegno di campionamento, calcolo e presentazione degli errori campionari*», Istat Indagine sulle strutture ed i comportamenti familiari, 11-27.

RUSSO A. (1986), «*Disegno di campionamento, calcolo e presentazione degli errori campionari*», Istat Indagine sulle letture e sugli altri aspetti dell'impiego del tempo libero nel 1984, Note e Relazioni n. 3, 13-27.

SADASIVAN, G. and AGGRAWAL, R. (1978), «*Optimum Points of Stratification in Bi-variate Populations*», Sankhia, 40, C, pp 84-97.

SETHI, V.K. (1963), «*A Note on Optimum Stratification of Population for Estimating the Population Means*», The Australian Journal of Statistics, 5, 20-23.

STATISTICS CANADA (1976), «*Methodology of the Canadian Labour Force Survey*», Household Surveys Development Division, Catalogue 71-526 occasional.

ZANI S. e SICURI S. (1977 a), «*Primi risultati di una stratificazione dei comuni dell'Emilia Romagna sulla base di indicatori socio-economici*» Istat, Commissione per gli studi statistici ed econometrici interessanti la programmazione economica, doc. n. 56.

ZANI S. e SICURI S. (1977 b), «*Stratificazione dei comuni dell'Emilia-Romagna con l'impiego del metodo gerarchico e del metodo non gerarchico*» Istat Commissione di studio per gli studi statistici ed econometrici interessati la programmazione economica, doc. n. 59.

ZANNELLA F. (1983), «*La progettazione del piano di campionamento per la seconda indagine sulle condizioni di salute della popolazione e sul ricorso ai servizi sanitari*» (Istat, Commissione di studio per le statistiche biologico-sanitarie).

ZANNELLA, F. (1984), «*La misura dell'errore delle stime nelle indagini campionarie multipurpose e l'utilizzazione di variabili ausiliarie nei procedimenti di stratificazione*», Atti delle XXXII Riunione Scientifica della S.I.S., Sorrento, 11-13 aprile, 321-327.

CAMPIONI DI NUMEROSITÀ DUE O TRE PER STRATO SELEZIONATI CON PROBABILITÀ VARIABILI: VALUTAZIONE EMPIRICA DI ALCUNE PROPRIETÀ DI STIME DI FREQUENZE ASSOLUTE.

di *Luigi Fabbris*

1. FINALITÀ DELLO STUDIO

Nella presente nota si commentano i risultati di una considerevole quantità di elaborazioni svolte su un archivio di dati comunali allo scopo di ottenere indicazioni sulle modalità di formazione di campioni di numerosità variabile da 1 a 3 per strato e sugli stimatori da adottare per formare campioni su più stadi per i quali si preveda al primo stadio la selezione con probabilità variabili di uno o più comuni per strato.

Gli stimatori sono descritti nel par. 2. Le elaborazioni di cui si riportano i risultati essenziali a partire dal par. 3 riguardano la selezione di 2 comuni per strato con tutte le tecniche di campionamento considerate e la selezione di 3 comuni per strato per 4 tecniche per le quali esiste una valida proposta di stimatore della varianza di campionamento. L'estensione delle considerazioni svolte per campioni di numerosità 3 a campioni di numerosità superiore, cambiando ciò che è logico cambiare, è automatica per le tecniche considerate.

Prima di presentare i risultati delle elaborazioni conviene chiarire il percorso logico che ha portato a concentrare l'attenzione sui temi e sul tipo di dati trattati.

1.1 La numerosità campionaria per strato

Per decenni, per le proprie indagini campionarie sulla popolazione, l'Istat ha seguito la via maestra del campionamento a più stadi stratificato con selezione al primo stadio di un comune per strato con probabilità proporzionali alla dimensione. Questa procedura è stata e continua ad essere utilizzata per l'indagine trimestrale sulle forze di lavoro (Istat, 1958, 1969; 1978) ed è stata adottata, parzialmente adattando il campione per la detta indagine, anche per molte indagini occasionali svolte dall'Istat sulla popolazione italiana.

Lasciando per ora sullo sfondo la necessità della coerenza tra gli obiettivi di ricerca e il disegno di campionamento — necessità che comporterebbe la formazione di campioni mirati, e quindi diversi, per ogni indagine — si discutono alcuni criteri per la scelta di un solo comune, due comuni, o più di due comuni per strato.

La selezione di un solo comune per strato in genere riduce la varianza fra unità che si esaminano. Purtroppo però non esiste uno stimatore corretto della varianza per un tale piano di campionamento. L'applicazione della tecnica «degli strati collassati», consistente nel raggruppare gli strati in insiemi di numerosità uguale o maggiore di due (Hansen, Hurwitz e Madow, 1953, vol. I: 9.15, 9.28 e vol. II: 9.5), come pure la stima della varianza con il criterio delle «differenze successive» (Kish, 1965: 8.6) portano ad una sovrastima della varianza di entità ignota e proporzionale allo scarto quadratico tra il valore stimato negli strati che si raggruppano e quello del «superstrato» che formano. È evidente che, se la distorsione è nulla per ogni strato, neppure la ulteriore stratificazione (di ogni superstrato in due strati più piccoli) contribuisce all'efficienza delle stime.

Hartley, Rao e Kiefer (1969), generalizzando un suggerimento di Hansen *et al.* (1953, vol. II: 5) propongono di stimare la varianza quando è stata estratta una sola unità per strato effettuando la regressione tra la stima in ognuno degli H strati formati e un vettore di variabili concomitanti strettamente correlate con il valore che si stima. Ciò riduce la distorsione nella stima proporzionalmente al valore del coefficiente di correlazione multipla tra il valore stimato e i regressori individuati. Hartley *et al.* (1969) non forniscono, tuttavia, lo stimatore appropriato per campioni di unità selezionati con probabilità variabili.

L'entità della distorsione nella stima conseguente al collassamento degli strati è proporzionale all'effetto della stratificazione più profonda.

La distorsione nella stima è comunque trascurabile per ogni finalità applicativa perché la stratificazione è generalmente poco efficace nel ridurre la variabilità delle stime dopo un certo numero di suddivisioni della popolazione statistica (Dalenius e Gurney, 1951; Dalenius, 1957: cap. 8; Hess e Srikantan, 1970; Cochran, 1977: 5A.8). Inoltre, il controllo sulla selezione si può ottenere con vari stratagemmi anche selezionando più comuni per strato (Goodman e Kish, 1950).

Shapiro e Olsen (1979) introducono un altro criterio per la scelta tra una o più di una unità di primo stadio. Essi si chiedono: se si utilizzano i risultati di una indagine per la verifica statistica di ipotesi di ricerca, quale tra le due alternative comporta un errore di secondo tipo più contenuto? Da varie prove su dati censuari, oltre all'ovvia indicazione che il miglior disegno è quello ad una sola unità primaria per strato se la distorsione nella stima della varianza è nulla, essi hanno trovato che la selezione di due unità per strato, con o senza reimmissione, dà stime significativamente più potenti (ossia, con errore di II tipo più contenuto) di quelle ricavabili con il disegno ad una sola unità per strato e non aggiustate per tener conto della distorsione. Gli Autori avvertono, tuttavia, che in altre situazioni di ricerca si possono ottenere indicazioni difformi da quelle dagli stessi descritte.

La scelta di un solo comune per strato sembra, in definitiva, non avere alcun elemento metodologico a favore. Indubbiamente, però, ne ha diversi a favore dal punto di vista pratico considerato che, a dispetto di tutte le indicazioni contrarie, una tale scelta è perpetuata dagli Istituti di statistica di vari Paesi.

Innanzitutto, la praticità si manifesta nella selezione, considerato che estrarre una unità primaria da ogni strato con probabilità variabile è considerevolmente più semplice che sceglierne due o più distinte.

Per indagini continue, nelle quali sia prevista la rotazione delle unità nel tempo, la sostituzione di una unità con un'altra scelta dallo stesso strato, meno numeroso e relativamente più omogeneo di un eventuale doppio strato, rende il numero di unità da selezionare dentro ogni unità di primo stadio, e conseguentemente la numerosità campionaria complessiva, più facilmente controllabile.

Le metodologia del campionamento con probabilità variabili per campioni di due unità per strato è sviluppata a sufficienza per ricavare soluzioni accurate sia per la formazione probabilistica di campioni complessi, sia per la stima corretta di statistiche lineari. Diverse tecniche per la formazione di campioni con probabilità variabili sono applicabili senza difficoltà anche per numerosità superiori a due per strato. Quasi tutte le tecniche sono, inoltre, programmabili per l'utilizzazione di strumenti di elaborazione automatica.

La scelta di un solo comune per strato è allora una soluzione di comodo, applicabile alle indagini continue, per le quali la soluzio-

ne operativamente più scorrevole è preferibile a ogni cambiamento «in corsa» indesiderato. Per queste indagini, inoltre, il calcolo dell'errore di campionamento si pone in occasione delle prime rilevazioni e di quando in quando, qualora si vogliano aggiornare i calcoli.

Per le indagini svolte *una tantum*, conviene, invece, che il campione sia, per quanto possibile, aderente agli obiettivi della ricerca e sia, dunque, progettato *ad hoc*, e che l'errore di campionamento sia calcolato di volta in volta. Per queste indagini la soluzione più semplice diventa allora quella di selezionare due comuni per strato.

Formando strati più ampi dai quali selezionare più di due comuni si ottengono stime più stabili della varianza delle statistiche che interessano. Bisogna, però ricordare che nelle indagini occasionali non interessano tanto le stime inerenti a singoli strati, quanto quelle di ambiti territoriali piuttosto vasti (talvolta la provincia, spesso la regione, più frequentemente la grande ripartizione geografica), ed è allora il campione dell'area vasta, ottenuto unendo i campioni degli strati che la compongono, a determinare l'efficienza e la stabilità delle stime.

La selezione di tre o più comuni per strato, riducendo — a parità di numerosità del campione globale — il numero di strati ottenibili con la selezione di due comuni per strato, è una scelta secondaria rispetto a quella della selezione di due unità di primo stadio in ogni strato, considerato che, a parità di altre condizioni, è dal numero di strati che dipende il controllo sulla selezione del campione.

Se si pensa, inoltre, alla rilevazione sistematica dell'errore dell'intervistatore mediante la tecnica dei campioni compenetranti (Mahalanobis, 1946), secondo la quale ad ogni intervistato è assegnato un campione casuale dell'intero strato e ogni strato è coperto da almeno due intervistatori, conviene creare strati territorialmente contenuti e 2 diventa il numero indicato di unità di primo stadio. D'altro canto, il numero di interviste che ogni rilevatore deve eseguire è fissato contrattualmente e il numero di comuni dello strato o del gruppo di strati che devono visitare gli intervistatori si può determinare in funzione di quel numero. Suggerimenti sulla relazione tra l'organizzazione del lavoro sul campo, l'efficienza attesa nelle stime e il numero di unità di primo stadio nella indagine sulle condizioni di salute della popolazione statunitense (*Health Interview Survey*) si trovano in Tadros, Moore e Chakrabarty (1982). Un progetto di nuova

indagine sulle forze di lavoro in Italia contenente anche la proposta di compenetrare le assegnazioni degli intervistatori è riportato in Fabbris (1981; 1983).

1.2 Indagini continue o occasionali

A questo punto è necessario condurre il ragionamento su due binari, quello delle indagini continue nel tempo (quali sono l'indagine trimestrale sulle forze di lavoro e quella mensile sui bilanci di famiglia) e quello delle indagini occasionali e variabili nel contenuto.

Per una indagine continua, sia la messa a punto del disegno, sia il calcolo dell'affidabilità delle principali stime, possono essere effettuati ogni tanto perché i cambiamenti nella distribuzione della popolazione sono impercettibili nel breve periodo e le modifiche nella struttura di rilevazione devono essere contenute nei limiti del possibile per una indagine avviata. Anche per questo motivo le indagini continue non rientrano se non marginalmente nelle considerazioni che si svolgono a favore della selezione di più comuni per strato, in quanto la valutazione dell'attendibilità delle stime può essere effettuata una volta ogni tanto. Non è un caso, tra l'altro, che spinte analoghe a quelle di cui si parla non abbiano indotto il Bureau of the Census degli USA a cambiare i criteri di selezione delle unità primarie per l'indagine sulle forze di lavoro *Current Population Survey*, che tuttora avvengono nel numero di uno per strato non autorappresentativo (National Commission..., 1979).

In un'indagine saltuaria ricorrono, invece, varie condizioni che rendono plausibili il soffermarsi sul disegno di rilevazione e il tentare un più effettivo sfruttamento dei dati dopo la raccolta. Tra le altre, la specificità dell'essere occasionali o a grande distanza di tempo dalla precedente, la disponibilità di tempo per predisporre accuratamente l'indagine, effettuare le elaborazioni, commentare i risultati e pubblicarli e, non ultime, le aspettative che si creano nei potenziali utenti dei risultati di tali indagini. Un punto che però consideriamo fisso nello sviluppare il nostro discorso è l'intelaiatura del campione: a due stadi, di cui almeno il primo sottoposto a campionamento, con stratificazione al primo stadio.

Per generalizzare la procedura di selezione del campione si propone di estrarre i comuni con probabilità variabili e ciò complica sia

la individuazione probabilistica delle unità campionarie, sia la stima delle statistiche di interesse e la valutazione delle loro proprietà.

Nel prosieguo si esaminano criteri comunemente utilizzati nelle indagini statistiche: il campionamento con o senza reinserimento delle unità estratte, la scelta sistematica, ma si effettuano anche incursioni in ambiti metodologici che, pur essendo meno frequenti nella prassi di ricerca, sono degni di attenzione in questa fase sperimentale: il calcolo di probabilità di selezione «ottime» con tecniche di analisi numerica, la selezione da sub-strati casuali, la selezione di campioni interi.

Le metodiche (par. 2.1) si valutano con riferimento alla *applicabilità*, vista sia come effettiva possibilità di implementare la tecnica nella realtà operativa dell'Istat, sia come maneggevolezza delle regole procedurali anche per persone in possesso di nozioni elementari di statistica (par. 5), e alle *proprietà tecniche degli stimatori* congruenti con la procedura di formazione del campione (par. 2.2), e in particolar modo alla efficienza degli stimatori (par. 3) e degli stimatori della loro varianza (par. 4).

Le proprietà tecniche degli stimatori sono valutate in base ad una cospicua serie di elaborazioni informatiche su variabili rilevate con il Censimento Generale della Popolazione e delle Abitazioni del 1981 in tre regioni, la Toscana, il Piemonte e la Calabria, assunte a rappresentanti delle regioni italiane. Naturalmente, la rappresentatività non è da intendersi nel significato che ha nel campionamento statistico, bensì per la tipicità delle loro caratteristiche: il Piemonte è una regione vasta, popolosa, composta da oltre 1100 comuni sotto i 20 mila abitanti, la Toscana ha una popolazione sparsa sul territorio e i comuni con meno di 20 mila residenti sono poco più di 250, la Calabria è la meno popolosa, ed in essa i Comuni sotto i 20 mila abitanti sono quasi 400.

Al fine di formare lo spazio dei possibili campioni si fa riferimento alla più recente stratificazione dei comuni utilizzata per l'indagine sulle forze di lavoro (luglio 1985), che prevede la suddivisione dei comuni in strati entro le province; da ogni strato, ogni tre anni, è sorteggiato un solo comune a rappresentare l'insieme dei comuni che compongono lo strato.

I comuni non autorappresentativi sono stati, inoltre, accoppiati così come fa il Servizio Istat addetto al calcolo della varianza di campionamento per le stime inerenti alle forze di lavoro. Il numero di strati

risultanti dall'abbinamento è 22 in Toscana, con 11,5 comuni in media per strato, 48 in Piemonte, con 24,6 comuni in media per strato, e 31 in Calabria, con 12,9 comuni in media per strato.

Per non rimanere ancorati nelle analisi ad una sola ipotesi di stratificazione, dentro ogni provincia, i comuni sono stati stratificati secondo l'ampiezza demografica. La regola di formazione degli strati è la seguente: dopo aver ordinato i comuni in base alla popolazione residente censita, sono stati formati tanti gruppi di comuni quanti sono i «superstrati» della stratificazione dianzi descritta, con l'avvertenza che i nuovi strati abbiano, dal punto di vista demografico, uguale peso. I comuni di dimensioni minori appartengono così a strati più numerosi; quelli più prossimi alla dimensione massima (20.000) sono raggruppati in strati di poche unità. La probabilità di selezionare un comune da questi ultimi strati è, dunque, maggiore di quella degli altri.

Nel formare campioni di terne di unità sono state mantenute le stesse stratificazioni della popolazione adottate per selezionare due comuni per strato.

Le due stratificazioni (d'ora in avanti dette anche «prima stratificazione» quella derivata dall'indagine sulle forze di lavoro e «seconda stratificazione» quella basata sulla dimensione dei comuni) rendono confrontabili stime provinciali. Ciò è conseguente alla congettura che in una indagine occasionale il livello territoriale minimo per il quale si producono stime pubblicabili sia la provincia. Nelle indagini occasionali o pluriennali finora svolte non sono state prodotte stime su scala territoriale inferiore alla regione.

Le valutazioni che si presentano sono finalizzate a fare la cernita delle procedure appropriate per formare un campione di comuni per una indagine sulla popolazione. Durante la gestazione dei giudizi che si stanno per presentare, in chi scrive, è maturata la convinzione che le «graduatorie di qualità» delle metodiche sono provvisorie e non univoche. La valutazione di altre proprietà o la valutazione delle stesse proprietà da angoli visuali diversi, possono rimescolare le posizioni acquisite dalle tecniche nelle graduatorie che si producono nel par. 6.

La presentazione dei risultati mira perciò a circoscrivere l'interesse di potenziali utenti alle tecniche che meglio rispondono a obiettivi generali, e ad identificare quelle da escludere senza patemi d'animo.

2. LE TECNICHE DI CAMPIONAMENTO E GLI STIMATORI CONSIDERATI

Le tecniche di campionamento considerate nella presente memoria (par. 2.1) sono quelle rimaste sul setaccio dopo varie prove descritte in Fabbris (1985). Alle tecniche si assegna un numero d'ordine che per semplicità le identificherà nelle elaborazioni e nelle analisi riportate in appresso.

Gli stimatori che si riportano nel par. 2.2 riguardano il numero di persone che possiedono un dato attributo (stime di livello), la loro varianza di campionamento, la stabilità degli stimatori dalla varianza in prove ripetute con campioni di numerosità 2 e 3.

2.1 Le tecniche di campionamento

Le procedure per le quali sono stati effettuati i calcoli della varianza campionaria e della varianza delle stime della varianza con campioni di numerosità due sono:

1. *la selezione con reinserimento delle unità estratte* (Hansen e Hurwitz, 1943). Le unità sono selezionate indipendentemente con probabilità proporzionali alla dimensione: il campione composto da due qualsiasi unità *i* e *j* appartenenti a un generico strato ha probabilità π_{ij} di essere selezionato:

$$\pi_{ij} = P_i P_j \qquad (i, j = 1, ..., N) \qquad (2.1)$$

dove $P_i = Z_i/Z$ denota la probabilità di selezione dell'unità *i* in una prova; Z_i è la dimensione demografica del comune *i*; inoltre

$$Z = \sum_{k}^{N} Z_k \qquad (2.2)$$

dove N è il numero di unità che compongono lo strato. In un campione di numerosità *n* la probabilità totale π_i di selezionare l'unità *i* è

$$\pi_i = nP_i; \qquad (2.3)$$

2. *il campionamento sistematico con stratificazione implicita dei comuni dentro lo strato* (Madow, 1949; Hartley, 1966). Dopo aver ordinato i comuni in base alla dimensione demografica, si formano N campioni distinti accoppiando i comuni secondo una tecnica piuttosto complessa da descrivere (cfr. Fabbris, 1985 App. A) che associa ad ogni campione s una probabilità tale che

$$\sum_{s}^{N} \pi_s = 1; \qquad (2.4)$$

dove π_s è la probabilità di selezionare il campione *s*;

3. *il campionamento sistematico con casualizzazione delle posizioni dei comuni nella lista* (Goodman e Kish, 1950; Horvitz e Thompson, 1952; Hartley e Rao, 1962). È il metodo tradizionale di selezione sistematica di unità con probabilità variabili: ad ogni unità si assegnano tanti numeri casuali (M_i) che, rapportati al totale dei numeri assegnati (M), eguagliano la probabilità P_i e si individuano le unità campionarie prendendone una ogni K (= M / n) a partire da un numero casuale compreso tra *1* e *K*. Ogni combinazione di unità è ammessa e la probabilità di selezionare l'unità *i* in tutti i possibili campioni è

$$\pi_i = 2P_i \qquad (i = 1, \ldots, N),$$

tuttavia è possibile che alcune coppie abbiano probabilità di inclusione nulla e questo è un ostacolo all'applicazione dello stimatore generale proposto nel par. 2, per cui si deve ricorrere all'artifizio proposto da Goodman e Kish (1950) per formare campioni aventi probabilità positiva, altrimenti si applica lo stimatore asintotico della varianza proposto da Hartley e Rao (1962);

4. *la selezione senza reimmissione di Brewer* (1963, 1975). La prima unità viene selezionata con probabilità

$$\pi_i' = \frac{2P_i\,(1-P_i)}{(1-2P_i)\,\lambda} \qquad (i = 1,\ldots,N) \qquad (2.5a)$$

dove:

$$\lambda = 1 + \sum_{k}^{N} \frac{P_k}{1-2P_k}, \quad 1-2P_k \qquad (2.5b)$$

e la seconda, senza reinserire quella estratta, con probabilità

$$\pi'_j = \frac{P_j}{1-P_i} \qquad (j = 1,..., N;\ j \neq i). \tag{2.6}$$

La probabilità totale di selezionare l'unità *i* è

$$\pi_i = 2P_i \qquad (i = 1,..., N)$$

se calcolata su tutti i possibili campioni. La probabilità di selezionare il campione formato dalle unità (i, j) è:

$$\pi_{ij} = \frac{2P_iP_j}{\lambda}\left[\frac{1}{1-2P_i} + \frac{1}{1-2P_j}\right] \qquad (i,j = 1,..., N;\ i \neq j). \tag{2.7}$$

Per n = 2 questa tecnica coincide con la proposta di Durbin e Sampford, descritta in Sampford (1967);

5. *il metodo del raggruppamento di Durbin* (1967: par. 4). Consiste nel raggruppare in G sottoinsiemi (G > 2), le unità appartenenti ad uno strato ed estrarre due unità dallo strato con reimmissione e con probabilità proporzionali alla dimensione. Se le unità selezionate provengono da gruppi diversi, le si accetta. Se provengono dallo stesso gruppo, la seconda viene reinserita e se ne estrae un'altra dallo stesso gruppo con probabilità

$$P'_j = P_j \,/\, \Sigma'_k P_k \tag{2.8}$$

dove Σ' denota la somma rispetto alle unità che appartengono al gruppo.

La probabilità di ottenere le unità *i* e *j* nell'ordine è $P_i\ P_j$ se le unità appartengono a gruppi diversi, e

$$\frac{P_iP_j}{\lambda'}\left[\frac{1}{1-2P'_i} + \frac{1}{1-2P'_j}\right], \tag{2.9a}$$

dove

$$\lambda' = 1 + \Sigma'_k \frac{P'_k}{1-2P'_k} \tag{2.9b}$$

se le unità provengono dallo stesso gruppo.

Durbin (1967) propone di formare quanti più gruppi possibili badando che $P_i \leq 0{,}5$ per ogni *i*. È evidente che raggruppamenti diversi danno campioni che variano in probabilità di selezione. Per rendere questa tecnica omogenea con un'altra nella quale prima della selezione si formano sottogruppi di comuni dentro lo strato (tecnica n. 7), si è deciso di adottare in ogni strato una sola classificazione dei comuni in $G = n$ gruppi con la logica descritta in relazione alla tecnica n. 7;

6. *il metodo iterativo di Fellegi* (1963), *Carroll e Hartley* (1964). Consiste nel selezionare la prima unità *i* con probabilità proporzionale a P_i e la seconda, senza reinserire quella estratta con probabilità P_j^* calcolata con una procedura iterativa finalizzata all'ottenimento di probabilità di inclusione quasi costanti ad ogni selezione. La probabilità di inclusione della unità *i*, in tutti i possibili campioni è, dunque

$$\pi_i \simeq 2P_i \qquad (i = 1,..., N),$$

e la probabilità di ottenere il campione (i, j)

$$\pi_{ij} = P_i \frac{P_j^*}{1 - P_i^*} + P_j \frac{P_i^*}{1 - P_j^*} \qquad (i, j = 1,...,N;\ i \neq j); \qquad (2.10)$$

7. *la procedura degli strati casuali di Rao, Hartley e Cochran* (1962). I sub-strati si formano suddividendo casualmente i comuni dello strato in $G = n$ gruppi, così che ogni sottogruppo è un campione casuale dei comuni dello strato. Da ogni sub-strato si estrae poi un comune campione con probabilità proporzionale alla sua dimensione demografica relativa a quella del substrato cui appartiene. Rao *et al.* (1962) indicano di formare substrati di numerosità N_g uguale

$$N_g = N / n \qquad (2.11)$$

se N / n dà un numero intero, e

$$N_g = (N + k) / n \qquad (2.12)$$

dove $k = 1,..., n-1$ se il rapporto N / n è un numero decimale.

Per condurre le analisi riportate in appresso i sub-strati sono stati formati con un algoritmo che tende a minimizzare la variabilità tra misure di grandezza Z_g ($g = 1,..., n$) dei gruppi da formare.

La probabilità di selezionare l'unità *i* in tutti gli N_1, N_2,..., N_g possibili campioni è

$$\pi_i = \frac{Z_i}{Z_{g\,(i)}} = \frac{P_i}{P_{g\,(i)}} \qquad (i = 1,..., N_{g\,(i)}) \qquad (2.13)$$

dove $Z_{g\,(i)}$ è la popolazione residente nel sub-strato che contiene anche l'unità i e $P_{g\,(i)} = Z_{g\,(i)}/Z$. Si osserva che se $Z_{g\,(i)} = Z/2$, $\pi_i = 2P_i$;

8. *l'estrazione senza reimmissione di Yates e Grundy* (1953). La procedura prevede la selezione senza reimmissione della prima unità *i* con probabilità P_i e della seconda con probabilità $P_j / (1\text{-}P_i)$. La probabilità annessa al campione (i, j) è

$$\pi_{ij} = P_iP_j \left[\frac{1}{1 - P_i} + \frac{1}{1 - P_j}\right] \qquad (i,j = 1,...,N;\ i \neq j), \qquad (2.14)$$

quella di inclusione dell'unità *i* non è strettamente proporzionale a Z_i su tutti i possibili campioni, essendo:

$$\pi_i = P_i + \sum_{k \neq i}^{N} \frac{P_iP_k}{1 - P_k} \qquad (i = 1,...,N); \qquad (2.15)$$

9. *la tecnica di rifiuto dei campioni non distinti di Yates e Grundy* (1953), che consiste nella selezione della prima unità *i* con probabilità P_i e, dopo la sua reimmissione, nella selezione della seconda unità *j* con probabilità P_j; se $i = j$ si rigetta l'intero campione e si ripete la selezione con probabilità immutate, finché $i \neq j$.

La probabilità di selezionare la coppia (i, j) è

$$\pi_{ij} = 2P_iP_j / (1 - \sum_{k}^{N} P_k^2) \qquad (1, j = 1,..., N;\ i \neq j) \qquad (2.16)$$

dove ΣP_k^2 è la probabilità che si selezioni una coppia di due unità uguali; la probabilità di selezionare la generica unità *i* in tutti i possibili campioni è

$$\pi_i = \frac{2P_i (1 - P_i)}{(1 - \sum_{k} P_k^2)} \qquad (i = 1,...,N); \qquad (2.17a)$$

la probabilità che l'unità *i* esca per prima in un campione di numerosità 2 è sempre

$$_{(1)}\pi_i = P_i / (1 - \sum_k^N P_k^2), \qquad (2.17b)$$

mentre la probabilità che essa dopo una qualsiasi delle altre unità, esclusi tutti i campioni non distinti, è

$$_{(2)}\pi_i = \frac{(1 - P_i) P_i}{1 - \sum_k P_k^2} ; \qquad (2.17c)$$

10. *la tecnica di determinazione delle probabilità di estrazione di campioni interi di Herzel* (1984), che si realizza assegnando agli N (N-1)/2 campioni distinti una probabilità che approssima la formula ideata da Brewer (1963; 1975) per la selezione senza reimmissione (tecnica 4). Per n = 2, la selezione del campione è effettuata assegnando ad ogni coppia distinta (i, j) la probabilità

$$\pi_{ij} = \pi_j\pi_j - \frac{\pi_i (1 - \pi_i) + \pi_j (1 - \pi_j)}{N - 2} + \\ + \frac{n - \sum_k^N \pi_k^2}{(N - 1) (N - 2)} \quad (i, j = 1,...,N; i \neq j) \qquad (2.18)$$

dove π_i può essere definita arbitrariamente purché

$$\pi_i \geq 0 \qquad (2.19a)$$

e

$$\sum_i^N \pi_i = n; \qquad (i = 1,..., N). \qquad (2.19b)$$

Per le nostre analisi si è fissato $\pi_i = 2P_i$.

11. *il campionamento casuale semplice* (senza reimmissione) che, come è ben noto, consiste nel selezionare in blocco due unità con probabilità uguali. La probabilità di estrarre un qualsiasi campione *s* è, dunque constante

$$\pi_s = \binom{N}{n}^{-1} \qquad (s = 1,..., \binom{N}{n}) \qquad (2.20)$$

La probabilità di estrazione della generica unità *i* nell'insieme dei possibili campioni è:

$$\pi_i = n / N \qquad (i = 1,..., N). \qquad (2.21)$$

La probabilità di selezionare l'unità *i* è altresì costante (1 / N) ad ogni estrazione.

Per campioni di numerosità superiore a 2 la probabilità di selezione si complica in alcuni casi in modo considerevole. Si è pensato di condurre le principali elaborazioni solo sulle procedure n. 1, 2, 4, 7 e 10 e solo per terne di comuni campione. Per il calcolo dell'efficienza si richiedono le specificazioni aggiuntive riportate nel seguito.

Tecnica 1. La probabilità di selezionare con reimmissione una qualsiasi delle N (N — 1) (N — 2)/6 terne (i, j, k) è data da

$$\pi_{ijk} = P_i P_j P_k \qquad (i, j, k = 1,..., N). \qquad (2.22)$$

Tecnica 2. Le terne campionarie che si possono formare con la tecnica proposta da Madow (1949) e Hartley (1966) sono *N* in uno strato di pari numerosità.

La probabilità π_s associata al campione *s* di numerosità *n* si calcola con l'algoritmo proposto dall'autore e schematizzato a fini di programmazione informatica in Fabbris (1985, App. A) per n = 2, e generalizzabile senza ulteriori difficoltà concettuali a *n* qualsiasi.

Tecnica 4. La probabilità associata ad una delle 3 coppie distinte di unità nelle quali è combinabile una terna (i, j, k) selezionata senza reimmissione con la procedura proposta da Brewer (1975) è data da

$$\pi_{ij} = \frac{2P_iP_j}{\lambda}\left[\frac{1}{1 - 3P_i} + \frac{1}{1 - 3P_j}\right] +$$

$$+ \sum_{k}^{N} \frac{P_k (1 - P_k)}{(1 - 3P_k)\,\lambda}\left\{\frac{P_i^* P_j^*}{\lambda^*}\left[\frac{1}{1 - 2P_i^*} + \frac{1}{1 - 2P_j^*}\right]\right\}$$

$$(i, j = 1,..., N; i \neq j) \qquad (2.23)$$

dove

$$P_i^* = P_i / (1 - P_k) \qquad (i = 1,..., N;\ i \neq k) \qquad (2.24a)$$

$$\lambda^* = \sum_{t \neq k}^{N} \frac{P_t^* (1 - P_t^*)}{1 - 2P_t^*} \qquad (2.24b)$$

$$\lambda = \sum_{t}^{N} \frac{P_t (1 - P_t)}{1 - 3P_t} \cdot \qquad (2.24c)$$

Non è stata, invece, trovata una formula soddisfacente per π_{ijk}.

Tecnica 7. La ripartizione della popolazione dello strato in *n* (n > 2) substrati segue la stessa logica esposta per n = 2. Vale la pena evidenziare che, per un dato *N*, l'applicazione della tecnica di Rao, Hartley e Cochran (1962) perde in efficacia con l'aumentare di *n*.

Per esempio, in Toscana, dove la numerosità media per strato è limitata, l'applicazione di questa procedura diventa piuttosto forzosa.

Il valore assunto da π_{ijk} è irrilevante per l'applicazione degli stimatori specifici proposti nel par. 2.2.

Tecnica 8. La probabilità associata ad una delle coppie di unità (i, j) che fanno parte di una terna selezionata con la tecnica senza reimmissione di Yates e Grundy (1953), è data da

$$\pi_{ij} = \sum_{j \neq i}^{N} \frac{P_i}{1 - P_i - P_j} + \sum_{j \neq i}^{N} \frac{P_j}{1 - P_i - P_j} + \qquad (2.25)$$

$$- (P_i + P_j) \left\{ \sum_{k}^{N} \frac{P_k}{1 - P_k} + \frac{P_i + P_j}{1 - P_i - P_j} \right\} \quad (i \neq j = 1,..., N).$$

Non è stata reperita nella letteratura una formula soddisfacente per π_{ijk}.

Tecnica 10. La logica dell'assegnazione della probabilità per la selezione di campioni interi proposta da Herzel (1984) è stata dianzi

esposta. Per n = 3 la probabilità congiunta della terna (i, j, k) è data da

$$\pi_{ijk} = \pi_i\pi_j\pi_k - \frac{b_{ij} + b_{jk} + b_{ik}}{N - 4} + \frac{b_i + b_j + b_k}{(N - 3)(N - 4)} + - \frac{b_o}{(N - 2)(N - 3)(N - 4)} \quad (i, j, k = 1,..., N) \quad (2.26)$$

dove

$$b_{ij} = \pi_i\pi_j (3 - \pi_i - \pi_j) - \pi_{ij} \quad (2.27a)$$

$$b_i = \pi_i (7 - 6\pi_i + 2\pi_i^2 - \sum_k^N \pi_k^2) \quad (2.27b)$$

$$b_o = 9 (3 - \sum_k^N \pi_k^2) - 2 (3 - \sum_k^N \pi_k^3) \quad (2.27c)$$

e π_{ij} è definita dalla (2.18).

Se per il calcolo di π_i si parte dalla dimensione dell'unità *i*,

$$\pi_i = 3P_i \qquad (i = 1,..., N).$$

2.2 Stimatori adottati

2.2.1 *Gli stimatori del totale*

Si consideri un insieme di *N* comuni in uno strato. La stima corretta del totale *Y* di una data variabile nella popolazione si ottiene sommando per ogni comune *i* incluso nel campione il rapporto tra il valore Y_i osservato nel comune e la probabilità π_i assegnata al comune nella fase di selezione (Horvitz e Thompson, 1952).

$$\hat{Y} = \sum_i^n \frac{Y_i}{\pi_i}. \quad (2.28)$$

Questo stimatore gode di importanti proprietà:

a) è il solo corretto nella classe degli stimatori che associano un peso unico alle unità statistiche selezionate (Horvitz e Thompson, 1952);

b) è il più efficiente nella classe degli stimatori lineari corretti del totale Y della popolazione (Roy e Chakravarti, 1960; Godambe, 1960);

c) se Y_i è legato da una relazione stocastica con π_i, con residui aventi media e covarianza nulla e varianza σ_i^2 data da

$$\sigma_i^2 = \sigma^2 Z_i^{2\gamma} \qquad 0{,}5 \leq \gamma \leq 1, \qquad (2.29)$$

la varianza di $\hat{Y}$ raggiunge il valore minimale tra tutti gli stimatori che hanno disegno «esatto», ossia: la selezione è senza reimmissione, $\pi_i \propto Z_i$, n è fisso (Godambe e Joshi, 1965);

d) se per ogni i ($i = 1,\ldots, N$) $Y_i = cZ_i$, dove c è una costante non nulla, e n è fisso, la varianza di $\hat{Y}$ è nulla. Questa proprietà è usualmente detta «dello stimatore quoziente»;

e) è lo stimatore di Bayes per tutte le distribuzioni a priori dell'estimando Y il cui valore atteso è proporzionale alla dimensione delle unità (Godambe e Joshi, 1965);

f) è l'unico «iper-ammissibile» secondo il criterio stabilito da Hanurav (1968) nella classe di tutti gli stimatori polinomiali corretti di Y.

Per l'estrazione casuale il blocco (tecnica n. 8) si adottano due stimatori, uno proposto da Raj (1956)

$$\hat{Y} = \sum_i^n t_i / n \qquad (2.30)$$

dove, per $n \geq 2$

$$t_i = \sum_i^{n-1} Y_i + Y_n \left(1 - \sum_i^{n-1} P_i\right) / P_n \qquad (2.31)$$

e uno proposto da Murthy (1957), detto «simmetrizzato», dato da

$$\hat{Y}'' = \sum_i^n \frac{P(s \mid i)}{P(s)} Y_i \qquad (2.32)$$

dove $P(s \mid i)$ denota la probabilità di selezionare il campione s condizionatamente alla selezione dell'unità i per prima e $P(s)$ è la probabilità non condizionata.

Per n = 2

$$\hat{Y}' = Y_i \frac{1 + P_i}{2P_i} + Y_j \frac{1 + P_i}{2P_j} \tag{2.33}$$

$$\hat{Y}'' = \frac{1}{2 - P_i - P_j} \left[(1 - P_j) \frac{Y_i}{P_i} + (1 - P_i) \frac{Y_j}{P_j} \right]. \tag{2.34}$$

Applicando la tecnica di Rao-Hartley-Cochran (tecnica n. 7), π_i è dato dalla formula (2.13).

2.2.2 *La varianza degli stimatori*

Per le tecniche denotate con i numeri 1, 4, 6, 9, 10, 11, se si assume che la numerosità campionaria sia fissa, la varianza di $\hat{Y}$, ricavata indipendentemente da Sen (1953) e Yates e Grundy (1953) (nel prosieguo richiamata con Sen-Yates-Grundy), è

$$V(\hat{Y}) = \sum_{i<j}^{N} \sum (\pi_i \pi_j - \pi_{ij}) \left[\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right]^2 \tag{2.35}$$

dove π_{ij} è la probabilità congiunta di estrarre i comuni *i* e *j*, mutabile, ovviamente, a seconda della tecnica di formazione del campione (cfr. par. 2.1).

Per la tecnica di selezione con reimmissione (n. 1) la formula si semplifica considerevolmente

$$V(\hat{Y}) = \frac{1}{n} \sum_{i>j}^{N} P_i P_j \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 \tag{2.36a}$$

$$= \frac{1}{n} \left(\sum_{i}^{N} \frac{Y_i^2}{P_i} - Y^2 \right) \tag{2.36b}$$

e ancor più con la selezione casuale semplice (tecnica n. 11):

$$V(\hat{Y}) = \frac{1}{n} \left(\sum_{i}^{N} NY^2 - Y^2 \right). \tag{2.37}$$

Per le rimanenti tecniche le varianze dello stimatore sono:

Tecnica 2 (Hartley, 1966):

$$V(\hat{Y}) \simeq \sum_{s}^{N} \pi_s (\hat{Y}_s - Y)^2 \tag{2.38}$$

dove π_s e $\hat{Y}_s$ denotano, rispettivamente, la probabilità e la stima associate al campione *s*.

Tecnica 3. Hartley e Rao (1962) suggeriscono di applicare una formula asintotica della varianza valida per qualsiasi *n*, approssimabile in:

$$V(\hat{Y}) = \sum_{i}^{N} \pi_i \left(1 - \frac{n-1}{n} \pi_i\right) \left(\frac{Y_i}{\pi_i} - \frac{Y}{n}\right)^2 \tag{2.39}$$

Tecnica 5. (Durbin, 1967):

$$V(\hat{Y}) = \sum_{i>j}^{N} \pi_{ij} w_{ij} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \tag{2.40}$$

dove

$$w_{ij} = \frac{\pi_i \pi_j}{\pi_{ij}} - 1 \tag{2.41}$$

in genere, e

$$w_{ij} = 1 \tag{2.42}$$

quando l'espressione (2.41) supera l'unità.

Tecnica 7. Per *n* qualsiasi (Rao, Hartley e Cochran, 1962):

$$V(\hat{Y}) = \left\{1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)}\right\} \left\{\sum_{i}^{N} \frac{Y_i^2}{nP_i} - \frac{Y^2}{n}\right\} \tag{2.43}$$

dove: $k = 0$ se N/n è un numero intero e $k = 1,\ldots, n-1$ se N/n è un numero decimale; *n* coincide con il numero di gruppi che si formano.

Tecnica 8. Nell'estrazione casuale senza reimmissione lo stimatore proposto da Raj (1956) ha varianza (Pathak, 1967a):

$$V(\hat{Y}) = \frac{1}{2n^2} \sum_{i>j}^{N} P_i P_j \left(1 + \sum_{r=2}^{n} \phi_{ij}(r-1)\right) \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 \tag{2.44}$$

dove ϕ_{ij} (r—1) è la probabilità di escludere *i* e *j* nelle prime (r—1) estrazioni, che, per n = 2 è

$$\phi_{ij}(1) = 1 - P_i - P_j. \tag{2.45}$$

Lo stimatore proposto da Murthy (1957) ha varianza (Pathak, 1967b):

$$V(\hat{Y}) = \sum_{i>j}^{N}\sum P_iP_j\left[1 - \Sigma_{s\epsilon(i,j)}\frac{P(s|i)\,P(s|j)}{P(s)}\right]\left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 \tag{2.46}$$

dove $\Sigma_{s\epsilon(i,j)}$ denota la sommatoria estesa a tutti i campioni che contengono le unità *i* e *j*. Per n = 2, la parte di formula entro parentesi quadra si riduce a

$$\frac{1 - P_i - P_j}{2 - P_i - P_j}. \tag{2.47}$$

La varianza dello stimatore della varianza di campionamento è genericamente espressa da

$$V(v(\hat{Y})) = \sum_s\left[v_s(\hat{Y})\right]^2\pi_s - \left[V(\hat{Y})\right]^2 \tag{2.48}$$

dove v_s denota lo stimatore della varianza del generico campione *s* (s = 1,..., N*), *N** è il numero di campioni che si possono formare nello strato, e π_s è la probabilità associata all'*s* - esimo campione:

$$\pi_s = \pi_{ij}$$

per il campione composto dalle unità *i* e *j* (n = 2), e

$$\pi_s = \pi_{ijk}$$

per la terna campionaria i, j, k (n = 3).

2.2.3 *Gli stimatori della varianza di campionamento*

Per le tecniche 1, 4, 6, 8, 10 e 11 ([1])

$$v_s(\hat{Y}) = \sum_{i>j}^{n}\sum\frac{n^2P_iP_j - \pi_{ij}}{\pi_{ij}}\left(\frac{Y_i}{nP_i} - \frac{Y_j}{nP_j}\right)^2 \tag{2.49}$$

che per la tecnica n. 1 si semplifica in:

$$v_s(\hat{Y}) = \frac{1}{n^2(n-1)} \sum_{i>j}^{n} \sum \left(\frac{Y_i}{nP_i} - \frac{Y_j}{nP_j} \right)^2 \qquad (2.50)$$

e per la n. 11 ulteriormente in

$$v_s(\hat{Y}) = \frac{N^2}{n^2(n-1)} \sum_{i>j}^{n} \sum (Y_i - Y_j)^2 \qquad (2.51)$$

Per le rimanenti tecniche gli stimatori della varianza dello stimatore sono:

Tecniche 2 e 3

$$v_s(\hat{Y}) = \frac{1}{n-1} \sum_{i>j}^{n} \left[1 - (\pi_i + \pi_j) + \sum_{k}^{N} \pi_k^2 / n \right] \left[\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right]^2 \qquad (2.52)$$

Per la tecnica 3, la probabilità del campione *s* - esimo, composto dalle unità *i* e *j* è data da:

$$\pi_s = \frac{(n-1)(N-2)^2 \pi_i \pi_j}{(N-1)(N-3)(n-\pi_i-\pi_j)} \left[1 - \frac{\sum_{k}^{N} \pi_k^2 - \pi_i^2 - \pi_j^2}{(n-\pi_i-\pi_j)^2} \right] \qquad (2.53)$$

Hanif e Brewer (1980) propongono, in alternativa alla (2.52), la formula approssimata

$$v(\hat{Y}) = \frac{n}{n-1} \left[1 - \frac{\sum_{j=1}^{N} \pi_j^{2\gamma}}{\sum_{j=1}^{N} \pi_j^{2\gamma} - 1} \right] \sum_{i=1}^{n} \left[\frac{Y_i}{\pi_i} - \frac{\hat{Y}}{n} \right]^2 \qquad (2.54)$$

dove $\gamma \simeq 0{,}75$.

Tecnica 5:

$$v_s(\hat{Y}) = w_{ij} \left[\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right]^2 \qquad (2.55)$$

dove w_{ij} è dato dalla formula (2.41) per i valori di w_{ij} che non eccedono 1 e dalla (2.42) in ogni altro caso.

Tecnica 7 (Rao, Hartley e Cochran, 1962):

$$v_s(\hat{Y}) = \frac{N^2 - k(n-k) - nN}{N^2(n-1) - k(n-k)} \sum_i^n P_{g(i)} \left[\frac{Y_i}{P_i} - \hat{Y} \right]^2 \qquad (2.56)$$

dove $P_{g(i)}$ denota la somma delle probabilità di selezione delle unità appartenenti al gruppo g che contiene l'unità campionaria i.

Tecnica 8:

Per lo stimatore proposto da Raj (1956)

$$v_s(\hat{Y}) = \frac{1}{n(n-1)} \sum_i^n (t_i - \hat{Y})^2, \qquad (2.57a)$$

dove t_i è definito dalla (2.31). Quando $n = 2$, $v_s(\hat{Y})$ si semplifica in

$$v_s(\hat{Y}) = (1 - P_i)^2 \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2. \qquad (2.57b)$$

La stima della varianza dello stimatore proposto da Murthy, per $n = 2$, è:

$$v_s(\hat{Y}) = (1 - P_i)(1 - P_j) \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2. \qquad (2.58)$$

Per il calcolo della (2.48), π_s è dato, per $n = 2$ e per ambedue gli stimatori, dalla (2.14).

2.2.4 *Stimatore basato sul quoziente*

Per ogni tecnica è stato applicato uno stimatore basato sul quoziente tra la somma dei valori osservati presso le unità campione e la somma delle probabilità di inclusione delle unità

$$r_s = \sum_i^n Y_i \,/\, \sum_i^n P_{(i)} \qquad (s = 1, \ldots, N^*) \qquad (2.59)$$

dove $P_{(i)}$ denota la probabilità di inclusione della unità i nell'ordine in cui è stata estratta. La scelta di $P_{(i)}$ invece di P_i vuole far risaltare le differenze tra tecniche di formazione del campione dipendenti dalla probabilità di inclusione delle unità.

Lo stimatore basato sul rapporto non è corretto, ossia $E(r) \neq Y$.

Per rendere confrontabile l'efficienza di questo stimatore con quelli specifici dianzi riportati si calcola la variabilità di *r* attorno a *Y*.

$$MSE(r) = \sum_s (r_s - Y)^2 \pi_s. \tag{2.60}$$

La stima di MSE (r) è, con una buona approssimazione, data da:

$$v_s(r) = \frac{N-2}{N(\Sigma P_{(i)})^2} \left[var(y) + r_s^2 \, var(p) - 2 r_s \, cov(yp) \right] \tag{2.61a}$$

che per $n = 2$ si riduce a

$$v_s(r) \simeq \frac{N-2}{N(P_i - P_j)^2} \left[(Y_i - Y_j)^2 + r_s^2 (P_i - P_j)^2 + \right.$$

$$\left. - 2 r_s (Y_i - Y_j)(P_i - P_j) \right] \tag{2.61b}$$

L'espressione si inserirà nella (2.48) per il calcolo della varianza dello stimatore della variabilità con π_s dato dalla probabilità di selezione del campione *s* composto dalla coppia di comuni *i* e *j* con le singole tecniche considerate.

Per stimare correttamente il livello della variabile *y* in *H* strati si adotta lo stimatore

$$\hat{Y} = \sum_h^H \hat{Y}_h \tag{2.62}$$

la cui varianza è data da:

$$V(\hat{Y}) = \sum_h^H V(\hat{Y}_h). \tag{2.63}$$

3. EFFICIENZA DEGLI STIMATORI

La sintesi regionale dei valori della varianza di 10 stime di totali, ottenute con campioni di due comuni per strato, suddivise per tecnica di campionamento, stimatore specifico e stimatore basato sul

quoziente è riportata nelle Tavv. 1 e 2 in Appendice, rispettivamente, per la stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro e per quella basata solo sulla dimensione demografica dei comuni.

Va sottolineato che, in relativo, si ottengono identiche indicazioni dai valori di efficienza espressi in rapporto alla varianza di una delle stime (nelle nostre analisi quella ottenuta con la tecnica del campionamento con reimmissione) e da quelli ottenibili facendo il rapporto con il valore stimato, considerato che, *mutatis mutandis*, gli stimatori di variabilità adottati sono tutti riferiti allo stesso valore stimato.

Il rapporto percentuale tra l'errore campionario delle stime e il

Prospetto 3.1 Coefficiente di variazione percentuale delle principali stime regionali ottenute con la tecnica di selezione con reimmissione *(Tecnica n. 1)* **e con la stratificazione derivata da quella dell'indagine sulle forze di lavoro** *(n = 2)*

VARIABILI DI CUI SI STIMA IL LIVELLO	TOSCANA		PIEMONTE		CALABRIA	
	M+F	F	M+F	F	M+F	F
Popolazione disoccupata	6,0	7,4	6,5	7,6	9,1	15,0
Popolazione occupata	1,1	2,5	0,8	1,5	2,0	4,8
Popol. in cond. professionali	1,1	2,4	0,7	1,4	1,6	4,4
Residenti con scuola obbligo	1,4	1,6	1,6	1,9	1,8	2,1
Residenti con laurea	8,3	8,9	5,6	6,7	7,9	9,3

Prospetto 3.2 Coefficiente di variazione percentuale delle principali stime regionali ottenute con la tecnica di selezione con reimmissione *(Tecnica n. 1)* **e con la stratificazione per dimensione demografica dei comuni** *(n = 2)*

VARIABILI DI CUI SI STIMA IL LIVELLO	TOSCANA		PIEMONTE		CALABRIA	
	M+F	F	M+F	F	M+F	F
Popolazione disoccupata	4,9	6,0	5,3	6,2	10,4	18,0
Popolazione occupata	0,9	2,0	0,6	1,2	1,9	4,8
Popol. in cond. professionali	0,9	2,0	0,6	1,2	1,8	5,2
Residenti con scuola obbligo	1,1	1,3	1,3	1,5	1,8	2,0
Residenti con laurea	6,7	7,3	4,6	5,5	7,0	8,4

valore della stima (detto anche *errore di campionamento relativo*) è riportato per la tecnica n. 1 nel Prospetto 3.1 in relazione alla prima stratificazione e nel Prospetto 3.2 a quella per dimensione.

Si osserva che:

— le stime regionali delle frequenze più elevate (la popolazione occupata in condizioni professionali e la popolazione con licenza della scuola dell'obbligo) sono considerevolmente efficienti, stante che l'errore campionario non raggiunge il 2% del valore stimato;

— le stime dei livelli più bassi sono, come è logico aspettarsi, meno efficienti di quelle appena considerate. I valori del coefficiente di variazione percentuale calcolati da Russo e Falorsi (1985) per alcune grandezze rilevate con l'indagine sulle forze di lavoro sono i seguenti:

		Toscana	Piemonte	Calabria
Popolazione disoccupata	M + F	14,0	8,6	26,0
	F	20,5	12,8	52,9
Forze di lavoro	M + F	1,3	0,7	2,2
	F	2,7	1,3	4,9

Si nota che l'indice relativo di variabilità dell'ammontare delle forze di lavoro calcolato da Russo e Falorsi è quasi uguale a quello della popolazione in condizioni professionali riportato nel Prospetto 3.1, mentre quello della popolazione disoccupata è considerevolmente diverso. Se si presta fede al valore dell'effetto del disegno di campionamento del livello della disoccupazione, che Russo e Falorsi (1985, Tav. 2) stimano in deff $\simeq$ 2, la differenza riscontrata tra le misure di efficienza non è facilmente spiegabile. Può darsi che la maggiore variabilità trovata da Russo e Falorsi sia imputabile al campionamento al secondo stadio;

— le stime dell'efficienza inerenti alla popolazione femminile non mostrano un andamento così differente da quello della popolazione nel complesso da meritare un trattamento a sé stante e nel prosieguo saranno commentate solo se divergono in modo percepibile dall'analogo valore inerente alla popolazione complessiva.

Un andamento affatto simile si rileva per campioni di numerosità 3. Infatti, essendo il numero di strati identico per ambedue le numerosità campionarie, l'unica differenza sta nel numero di comuni campione globalmente selezionati. Ciò vale a dire che, per passare da un valore di efficienza per $n = 2$ ad uno con n qualsiasi per strato, *ceteris paribus*, basta moltiplicare il valore derivante dal campione di numerosità 2 per $\sqrt{2/n}$. Partendo dai valori riportati nelle Tabb. 3.1 e 3.2, l'errore relativo di campionamento per $n = 3$ è ottenuto moltiplicandoli per 0,82 ($= \sqrt{2/3}$) e per $n = 1$ è 1,41 ($= \sqrt{2}$). Il cambio di numerosità incide, però, in vario modo sulle stime ottenute con piani di formazione del campione diversi da quello con reimmissione.

Prima di iniziare l'analisi delle differenze tra tecniche vale la pena commentare un altro elemento «sovrastrutturale»: l'efficacia dei due tipi di stratificazione sperimentati nel contenere la variabilità delle stime. Posta uguale a 100 la varianza di campionamento ottenuta da campioni selezionati con reimmissione (Tecnica n. 1) stratificati come per l'indagine sulle forze di lavoro, i valori della varianza della stessa stima con la stratificazione basata sulla dimensione demografica dei comuni (Prospetto 3.3) sono in alcuni casi inferiori e in altri superiori a 100.

La media geometrica dei valori calcolati con le 10 variabili è 99,0 per la Toscana, 125,6 per il Piemonte e 93,1 per la Calabria. La media globale (101,5) è superiore al valore fissato come riferimento. Anche se le grandezze esaminate non sono da considerarsi un campione rappresentativo di tutti i contenuti che possono interessare nelle indagini sulla popolazione né le due regioni dell'Italia intera (2), i valori calcolati indicano che, nella gran massa, possiamo aspettarci maggiore efficacia nel contenere la variabilità delle stime dalla stratificazione basata sulla dimensione, rispetto ad altre (apparentemente) più mirate forme di stratificazione.

Nel predisporre una indagine occasionale sulla popolazione italiana, per la quale si preveda di selezionare un sottomultiplo del numero di comuni attualmente selezionati per l'indagine sulle forze di lavoro (circa 1900, sovracampionamenti compresi), la dimensione degli strati diventa superiore a quella media esaminata e si prospetta una situazione che è più prossima a quella del Piemonte che a quella della Toscana.

Prospetto 3.3 Rapporto percentuale tra la varianza delle stime ottenute con campioni di numerosità 2 (a) selezionati con reimmissione dagli strati derivati dalla stratificazione adottata per l'indagine sulle forze di lavoro e quelli formati in base alla dimensione demografica dei comuni.

	TOSCANA		PIEMONTE		CALABRIA	
	M+F	F	M+F	F	M+F	F
Popolazione disoccupata	97,8	99,4	162,3	167,9	75,8	69,3
Popolazione occupata	94,0	96,5	110,5	106,3	107,0	97,5
Popol. in cond. professionali	94,4	97,0	114,4	108,6	78,2	70,3
Residenti con scuola obbligo	95,4	97,5	195,4	190,9	103,3	103,6
Residenti con laurea	107,9	111,3	71,5	92,1	125,6	120,5

(a) L'efficienza complessiva dello stimatore non varia in ragione della numerosità del campione selezionato in ogni strato. - (b) Nel calcolare le stime con la stratificazione derivata da quella per l'indagine sulle forze di lavoro sono stati esclusi due strati con valori anomali. L'efficienza della stratificazione basata sulla dimensione demografica è, pertanto, leggermente superiore a quella qui riportata.

Oltre che elementare nell'applicazione, la stratificazione basata sulla dimensione comunale è allora anche efficiente se si adotta uno dei tradizionali disegni di campionamento a più stadi e selezione al primo stadio con probabilità proporzionali alla dimensione.

Per quanto attiene alle singole tecniche, i risultati delle quali sono ulteriormente riepilogati nei Prospetti 3.4 e 3.5, si riportano in appresso alcune sintetiche considerazioni.

Tecnica 2. Anche se in diversi casi viene superata in efficienza da altre tecniche, la selezione sistematica da strati ordinati è globalmente la più efficiente, sia per $n = 2$, sia per $n = 3$. Se si adotta la formula della varianza proposta da Hartley (1966), dà stime regionali regolarmente più efficienti di quelle dei campioni ottenuti con reimmissione.

Per $n = 2$, l'efficienza dello stimatore basato sul quoziente è monotona con quella dello stimatore specifico ([3]).

Prospetto 3.4 Media geometrica dei valori di efficienza riportati nelle Tabb. A.1 e A.2 per tecnica di formazione dei campioni di 2 comuni, per regione, stimatore e tecnica di stratificazione dei comuni *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente; I: stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro; II: stratificazione per dimensione demografica dei comuni; NC: Non Calcolabile)*

TECNICA DI FORMAZIONE DEL CAMPIONE	TOSCANA				PIEMONTE				CALABRIA			
	I		II		I		II		I		II	
	St. A	St. B	St. A	St. B	St. A	St. B	St. A	St. B	St. A	St. B	St. A	St. B
1	100	94,0	100	100,5	100	97,8	100	99,8	100	92,8	100	100,0
2	127,9	125,4	119,5	118,9	111,6	106,2	107,7	104,9	112,4	95,4	115,8	115,7
3	111,8	106,6	112,8	NC	106,1	103,2	107,3	109,3	110,4	102,8	112,1	114,8
4	114,6	63,4	115,7	68,2	107,8	74,2	108,5	59,4	112,9	80,6	114,8	93,9
6	NC	NC	108,8	108,6	NC	NC	98,9	98,8	98,1	91,0	112,2	112,9
7	109,4	101,8	114,5	113,0	103,7	102,3	108,0	109,9	107,6	99,5	113,6	120,1
8 Raj	112,4	80,1	113,2	73,3	107,2	100,0	108,2	63,2	111,2	95,5	112,1	95,6
Murthy	114,7		116,2		108,0		109,6		113,0		114,8	
9	81,1	103,8	100,1	113,0	71,0	103,3	103,7	103,7	98,2	100,9	114,1	114,5
10	112,3	107,7	113,8	114,5	96,8	89,3	109,5	109,1	111,8	103,5	114,8	114,7
11	3,3	6,2	13,9	16,5	1,7	3,1	37,5	44,5	8,9	9,6	75,6	87,0

Prospetto 3.5 Media geometrica dei valori di efficienza ottenuti applicando stimatori specifici *(Tab. A.3)* **per tecnica di formazione di campioni di 3 comuni, per regione e tecnica di stratificazione dei comuni** *(I: stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro; II: stratificazione per dimensione demografica dei comuni)*

TECNICA DI FORMAZIONE DEL CAMPIONE	TOSCANA		PIEMONTE		CALABRIA	
	I	II	I	II	I	II
1	100	99,0	100	125,6	100	93,1
2	129,0	153,6	128,5	153,3	137,4	130,0
4	134,4	136,0	117,1	151,9	130,1	125,6
7	120,9	125,7	107,5	148,8	115,9	119,4
10	132,1	131,2	107,2	151,9	128,6	125,6

Tecnica 3. Applicata solo per n = 2, per le difficoltà di reperire uno stimatore appropriato, la selezione sistematica da liste casuali dà stime generalmente più efficienti di quelle ottenute con la tecnica standard della reimmissione ma, sembra quasi un'ironia, i migliori risultati si ottengono quando i comuni sono stratificati implicitamente dentro gli strati.

Tecnica 4. L'estrazione in blocco con le probabilità di selezione determinate da Brewer (1963; 1975) dà stime più efficienti di quelle del-

campionamento con reimmissione, con valori di efficienza regolarmente crescenti con l'aumentare della frazione sondata dentro gli strati.

L'efficienza del campionamento in blocco è talvolta superiore a quella del campionamento sistematico con la formula della probabilità di Hartley (1966). Come questo, comunque, dà stime maggiormente efficienti quando i comuni sono raggruppati in insiemi omogenei per dimensione demografica.

La selezione senza reimmissione secondo Brewer è, però, nettamente inefficiente se si adotta uno stimatore basato sul quoziente tra la somma dei valori osservati e la somma delle probabilità di selezione condizionate all'ordine di inclusione nel campione.

Tecnica 5. Le condizioni per l'applicabilità della tecnica del raggruppamento proposta da Durbin (1967) sono frequentemente violate nel calcolo delle stime sia per la Toscana, sia per il Piemonte. Per una efficace applicazione di questa tecnica va studiata una stratificazione *ad hoc* delle unità da sottoporre a campionamento. A giudizio di chi scrive, questa procedura si autoesclude da ogni considerazione nella ricerca su dati reali perché macchinosa, applicabile solo se la dimensione di ogni comune nello strato è inferiore a $1/n$ nei sub-strati costruiti. In Toscana la procedura di Durbin non è applicabile in 7 strati su 22 e in Piemonte in altrettanti su 46 con la stratificazione derivata da quella impiegata per svolgere l'indagine sulle forze di lavoro.

Per valutare la convenienza della tecnica dal punto di vista dell'efficienza delle stime si riportano i coefficienti di variazione percentuali di due variabili ottenuti per n = 2 per le sole province nelle quali la tecnica è correttamente applicabile, e li si confronta con gli analoghi valori risultanti dall'applicazione della tecnica sistematica di Madow-Hartley e di quella senza reimmissione di Brewer (1963).

LAUREATI		TECNICHE DI FORMAZIONE DEL CAMPIONE		
		n. 5	n. 2	n. 4
46. Lucca	M + F	28,0	28,6	24,7
	F	28,3	28,1	25,3
48. Firenze	M + F	27,9	20,8	26,9
	F	28,0	20,9	27,0
50. Pisa	M + F	16,6	13,5	17,0
	F	17,8	13,6	18,2
51. Arezzo	M + F	24,2	21,5	21,0
	F	32,0	30,6	29,0
3. Novara	M + F	11,4	10,6	10,5
	F	14,0	13,4	12,8
6. Alessandria	M + F	10,9	12,0	11,9
	F	15,3	16,7	16,6

Si percepisce immediatamente che la tecnica n. 10 non dà neppure vantaggi costanti in termini di efficienza rispetto alle altre due.

Tecnica 6. Le probabilità ricavate con il metodo iterativo di Fellegi (1963) sono applicate solo a campioni di coppie di unità. L'algoritmo di stima delle probabilità non converge solo in uno strato in Toscana e in un altro in Piemonte.

I campioni che si ricavano sono sempre più efficienti di quelli ottenuti reimmettendo le unità estratte (cfr. anche Rao, 1963) e danno risultati che non si discostano granché da quelli dei campioni sistematici o con reimmissione. Risultati consonanti sono ottenuti, per quanto concerne le tecniche n. 4 e n. 6, da Rao e Bayless (1969) su 20 popolazioni naturali di piccole dimensioni ($N \leq 35$). I dati che seguono, analoghi a quelli prodotti per la tecnica n. 5, possono servire per analisi impressionistiche dell'efficienza differenziale delle tecniche messe a confronto.

		TECNICHE DI FORMAZIONE DEL CAMPIONE		
		N. 6	N. 2	N. 4
Laureati in Toscana	M + F	7,6	6,8	7,7
	F	8,0	7,3	8,0
Laureati in Piemonte	M + F	4,9	5,3	5,2
	F	5,9	6,5	6,3

Brewer e Hanif (1983: Tab. 3.2) trovano che la varianza dello stimatore di Horvitz-Thompson cui sono applicate le probabilità di estrazione ricavate con l'algoritmo di stima proposto da Fellegi (1963) è prossima alla varianza minima descritta dagli stessi Brewer e Hanif (1969).

Tecnica 7. La procedura degli strati casuali di Rao-Hartley-Cochran dà stime la cui variabilità è desumibile da quella del campionamento bernoulliano, considerato che (Rao *et al.*, 1962)

$$V(\hat{Y})_{(7)} = V(\hat{Y})_{(1)} \left\{ 1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)} \right\} \qquad (3.1)$$

dove $0 \leq k < n$ (cfr. formule 2.36 e 2.43); e il deponente di $V(\hat{Y})$ indica la tecnica di campionamento cui si applica. Il guadagno in efficienza sul campionamento con reimmissione è dunque pressoché

costante a livello regionale ed è tanto più cospicuo quanto più è elevata la frazione di campionamento. In Toscana, dove la frazione di comuni campionati è mediamente il 17%, il guadagno in efficienza con la stratificazione derivata dal campionamento per l'indagine sulle forze di lavoro per $n = 2$ e $n = 3$ è, rispettivamente, attorno al 9% e al 21%; in Piemonte, dove la frazione che si ipotizza di sondare è meno della metà (8% per $n = 2$ e 12% per $n = 3$), il guadagno è, rispettivamente, del 4% e dell'8%; in Calabria, dove la frazione è intermedia, il guadagno è dell'8% e del 16%.

Applicata alla stratificazione per dimensione dei comuni, la tecnica di Rao-Hartley-Cochran dà risultati generalmente migliori dell'altra stratificazione sperimentata.

Nelle varie applicazioni la tecnica di cui si tratta è quasi sempre meno efficiente sia del campionamento sistematico con stratificazione implicita, sia del campionamento senza reimmissione con le probabilità date da Brewer (1963; 1975). Questo risultato è sorprendentemente diverso da quelli ottenuti da Rao e Bayless (1969) e Bayless e Rao (1970), i quali trovano indicazioni favorevoli alla tecnica di Rao-Hartley-Cochran soprattutto sovraimponendo ai dati un modello di superpopolazione. Se i nostri calcoli sono corretti, il modello non è valido nella realtà che si esamina.

Stratificando per dimensione e formando strati di numerosità sufficientemente larga si ottengono praticamente gli stessi risultati con le tre tecniche considerate (n. 2, 4 e 7).

Con la stratificazione derivata dall'indagine sulle forze di lavoro, si verifica frequentemente la condizione che la correlazione intrastrato sia positiva, oppure che sia verificata la disuguaglianza

$$\sum_i^N P^2 \left(\frac{Y_i}{P_i} - Y\right)^2 > \frac{1}{N} \sum_i^N P_i \left(\frac{Y_i}{P_i} - Y\right)^2 \qquad (3.2)$$

affinchè la varianza asintotica proposta da Hartley e Rao (1962) sia inferiore a quella proposta da Rao *et al.* (1962) per la selezione di substrati casualmente determinati. Stranamente, ciò non si verifica quando gli strati sono formati in base alla dimensione demografica. Sul risultato può incidere l'approssimazione della formula asintotica per la quale si assume che N sia molto grande.

Tecnica 8. Sui campioni selezionati con la tecnica di estrazione senza reimmissione di Yates e Grundy (1953) lo stimatore proposto

da Murthy (1957) è sistematicamente più efficiente di quello prima proposto da Raj (1956), e a quello si fa, dunque, riferimento per confronti con le altre tecniche.

Va detto innanzitutto che la maggiore efficienza del campionamento senza reimmissione sul campionamento con reimmissione è contenuta: 15-16% in Toscana, 8-10% in Piemonte e 13-15% in Calabria per campioni di numerosità 2, i soli ai quali è stata applicata la selezione senza reimmissione di Yates-Grundy.

Il confronto più interessante si effettua per $n = 2$ con l'altra tecnica di selezione senza reimmissione (Brewer, 1963), per la quale si applica lo stimatore di Horvitz-Thompson per il totale e quello di Sen-Yates-Grundy per la varianza. Seppure di poco (1-2%), lo stimatore di Murthy risulta più efficiente di quello comunemente applicato, come trovano anche Rao e Bayless (1969). Bayless e Rao (1970) giungono a conclusioni analoghe anche per $n = 3$ e $n = 4$. Rao (1966), applicando ai suoi dati un modello di superpopolazione, trova una relazione altalenante nell'efficienza tra le due tecniche.

Va, inoltre, ricordato che, diversamente da quello di Sen-Yates-Grundy, lo stimatore della varianza del totale proposto da Murthy non è mai negativo.

Lo stimatore di Murthy è anche più efficiente dello stimatore applicato da Rao *et al.* (1962) a campioni formati con la tecnica proposta contestualmente allo stimatore. I nostri risultati concordano con quelli ottenuti da Rao e Bayless (1969) sulle già citate 20 popolazioni naturali e su 7 popolazioni artificiali per $n = 2$ e su altre 14 popolazioni naturali ($10 \leq N \leq 20$) per $n = 3$. Questo risultato è supportabile anche analiticamente: Pathak (1966) dimostra che, sotto condizioni abbastanza generali, la varianza dello stimatore di Murthy è inferiore a quello dello stimatore di Rao-Hartley-Cochran.

La selezione senza reimmissione con stimatori proposti da Murthy (1957) per il totale e da Pathak (1966) per la varianza è meno efficiente della selezione sistematica con stimatore proposto da Hartley (1966) nella maggior parte delle applicazioni riportate nelle Tavv. 1 e 2. Invece, in Piemonte con la stratificazione per dimensione dei comuni e in Calabria con quella derivata dall'indagine sulle forze di lavoro, la tecnica di cui si tratta diventa più conveniente di quella sistematica. Anche con la tecnica n. 8 la stratificazione per dimensione si dimostra più efficiente.

Tecnica 9. Alla tecnica di rifiuto di campioni non distinti proposta da Yates e Grundy (1953) sono stati applicati gli stimatori di Horvitz-Thompson per il totale e di Sen-Yates-Grundy per la varianza. La varianza delle stime che si ricavano è spesso superiore anche a quella ottenuta reinserendo volta a volta le unità estratte.

Tra tutte le tecniche considerate, questa è l'unica per la quale l'efficienza dello stimatore basato sul quoziente tra la somma dei valori rilevati e la somma delle probabilità di inclusione supera quasi sempre quella dello stimatore convenzionale.

Tecnica 10. La selezione di campioni interi probabilizzati secondo la proposta di Herzel (1984) e la susseguente proposta di eliminazione delle probabilità negative, producono stime generalmente più efficienti dell'applicazione dello stimatore di Murthy (1957) a campioni selezionati con reimmissione.

Per costruzione, l'efficienza è all'incirca uguale a quella inerente alla tecnica di selezione in blocco secondo Brewer (1975), di cui la formulazione proposta da Herzel è *latu senso* una generalizzazione approssimata. I risultati sono più evidenti per campioni composti da terne di unità. Questa tecnica è maggiormente conveniente dal punto di vista della variabilità delle stime quando la selezione avviene da strati più «regolari», quali sono quelli costruiti in base alla omogeneità della dimensione demografica dei comuni.

Tecnica 11. Il campionamento casuale semplice è stato provato solo per dare un significato particolare al campionamento con probabilità variabili. Dai dati riportati nei Prospetti 3.4 e 3.5 è chiaro che la selezione con probabilità variabili, qualora, come nel caso di cui si tratta, siano anche moderatamente proporzionali alla variabile oggetto di rilevazione, esercita un controllo nella selezione che rende le stime efficienti in modo e in quantità non inferiori a una stretta stratificazione per dimensione. A questo proposito, Raj (1958), ritenendo ipotizzabile un modello di superpopolazione espresso da:

$$Y_{ih} = \alpha + \beta Z_i + \epsilon_{ih} \qquad (3.3)$$

dove Z_i è la dimensione dell'unità *i* e *h* è lo strato cui appartiene, dimostra che, sotto condizioni verosimili, lo stimatore derivante dalla selezione con probabilità variabili è sempre più preciso di quello basato sulla stratificazione proporzionale, ma meno di quello «ottimale» secondo Neyman-Tschuprow.

4. STABILITÀ DEGLI STIMATORI DELLA VARIANZA

L'efficienza degli stimatori della varianza annessi alle tecniche di selezione adottate è riportata nelle Tavv. 4 e 5 per n = 2, e 6 per N = 3, con riferimento alla varianza dello stimatore della varianza (del totale) nel campionamento con reimmissione; la massa di risultati è riepilogata nei Prospetti 4.3 e 4.4. Il rapporto percentuale tra l'errore campionario della varianza e la varianza stessa è presentato nei Prospetti 4.1 e 4.2 per tutte le variabili considerate.

Prospetto 4.1 Coefficiente di variazione percentuale delle principali stime regionali della varianza ottenute con campioni di numerosità 2 con reimmissione *(Tecnica n. 1)*, **per tipo di stratificazione adottato.**

STRATIFICAZIONE/ VARIABILE	TOSCANA		PIEMONTE		CALABRIA	
	M+F	F	M+F	F	M+F	F
Stratificazione indagine forze di lavoro						
Popolazione disoccupata	39,7	47,9	37,7	43,8	48,6	57,9
Popolazione occupata	37,9	42,5	35,0	35,0	33,8	31,4
Popol. in cond. professionali	40,2	43,1	35,2	34,2	40,0	30,6
Residenti con scuola obbligo	36,1	38,9	38,8	38,3	32,0	31,6
Residenti con laurea	98,7	85,5	40,8	39,5	52,7	49,3
Stratificazione basata sulla dimensione dei comuni						
Popolazione disoccupata	36,5	47,7	30,3	36,8	39,0	47,4
Popolazione occupata	35,1	38,8	26,1	25,0	23,6	21,9
Popol. in cond. professionali	36,1	39,0	26,5	24,8	27,6	21,4
Residenti con scuola obbligo	40,0	41,2	27,3	25,4	30,4	30,5
Residenti con laurea	79,2	71,3	79,9	62,4	33,8	32,0

Prospetto 4.2 Coefficiente di variazione percentuale delle principali stime regionali della varianza ottenute con campioni di numerosità 3 con reimmissione *(Tecnica n. 1)*, **per tipo di stratificazione adottato.**

STRATIFICAZIONE VARIABILE	TOSCANA		PIEMONTE		CALABRIA	
	M+F	F	M+F	F	M+F	F
Stratificazione indagine forze lavoro						
Popolazione disoccupata	29,0	35,5	26,8	32,5	36,4	44,0
Popolazione occupata	27,5	31,0	25,9	25,6	24,6	22,1
Popol. in cond. professionali	40,6	43,0	26,0	24,9	30,1	21,6
Residenti con scuola obbligo	25,6	27,8	27,8	27,5	22,6	22,5
Residenti con laurea	74,3	63,8	30,1	28,8	38,7	35,7
Stratificazione basata sulla dimensione dei comuni						
Popolazione disoccupata	26,9	35,7	22,4	28,1	28,5	35,6
Popolazione occupata	25,2	27,8	18,9	19,3	23,2	26,5
Popol. in condizioni professionali	26,2	28,3	19,3	17,8	21,1	20,1
Residenti con scuola obbligo	29,7	30,1	19,9	18,2	21,3	21,3
Residenti con laurea	58,2	52,5	61,6	48,1	27,3	30,7

Prospetto 4.3 Media geometrica dei valori di efficienza riportati nelle Tabb. A.4 e A.5 per tecnica di formazione di campioni di coppie di comuni, per regione, stimatore e tecnica di stratificazione dei comuni *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente; I: stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro; II: stratificazione per dimensione demografica dei comuni; NC: Non Calcolabile)*

TECNICA DI FORMAZIONE DEL CAMPIONE	TOSCANA				PIEMONTE				CALABRIA			
	I		II		I		II		I		II	
	St. A	St. B	St. A	St. B	St. A	St. B	St. A	St. B	St. A	St. B	St. A	St. B
1	100	195,6	100	192,5	100	216,3	100	146,8	100	184,3	100	151,3
2	111,7	137,5	133,9	179,3	96,6	268,0	117,9	144,4	122,7	186,8	107,4	111,0
3	134,5	188,1	136,8	179,9	110,1	216,2	119,1	140,2	127,1	182,2	137,4	141,5
4	129,9	88,9	134,8	82,6	108,1	146,4	119,0	46,4	121,8	125,4	137,2	109,3
6 (a)	203,4	124,8	246,9	156,6	187,8	108,0	261,6	121,4	183,7	137,5	247,5	138,0
7	118,3	195,5	137,3	181,9	104,2	214,0	96,3	132,6	120,8	176,2	107,9	137,1
8 Murthy	249,5	129,5	248,3	95,7	242,2	215,4	281,4	68,1	241,6	165,6	235,3	113,5
9	153,1	184,2	213,7	176,2	110,7	110,7	280,0	137,7	191,5	178,0	239,9	141,5
10	50,1	179,0	114,6	176,7	NC	NC	NC	134,4	27,6	173,2	136,4	141,4
11	0,0	0,0	1,6	2,1	0,0	0,0	10,9	13,6	0,0	0,0	48,1	73,6

(a) La varianza della varianza è stata calcolata escludendo lo strato per il quale l'algoritmo di calcolo delle probabilità di selezione non converge. La stabilità è, dunque, leggermente sovrastimata.

Prospetto 4.4 Media geometrica dei valori di efficienza ottenuti applicando stimatori della varianza *(Tab. A.6)* **per tecnica di formazione di campioni di 3 comuni, per regione e tecnica di stratificazione dei comuni** *(I: stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro; II: stratificazione per dimensione demografica dei comuni; NC: Non Calcolabile)*

TECNICA DI FORMAZIONE DEL CAMPIONE	TOSCANA		PIEMONTE		CALABRIA	
	I	II	I	II	I	II
1	100	112,7	100	198,2	100	114,0
2	163,0	166,1	109,6	224,5	124,8	176,7
7	150,4	215,9	128,7	352,0	136,7	174,4
10	NC	191,7	NC	NC	NC	212,5

Si osserva che:

— la stabilità della varianza campionaria è piuttosto bassa considerando che la radice quadrata della sua varianza è, per $n = 2$, in media sull'insieme delle variabili considerate, il 51% del suo valore corretto in Toscana, il 38% in Piemonte e il 36% in Calabria e che per $n = 3$ la varianza si riduce a 1/4 del valore in assoluto, mentre il coefficiente di variazione scende solo al 40% in Toscana e al 28% in Piemonte e in Calabria;

— il coefficiente di variazione indica che i campioni selezionati da strati formati in base alla dimensione demografica dei comuni godono di una maggiore stabilità complessiva;

— lo stimatore basato sul quoziente tra la somma dei valori rilevati e la somma delle probabilità di estrazione stabilizza le stime della varianza rispetto agli stimatori specifici o asintotici e livella le differenze in stabilità delle varie tecniche.

Gli stimatori basati sul quoziente tendono a concentrarsi attorno alla media più degli stimatori specifici in ragione della correlazione esistente tra la stima della statistica e la probabilità di selezione delle unità. Se, cioè, si ipotizza una relazione di proporzionalità tra le due variabili (v., tra gli altri: Rao, 1967)

$$Y_i = \beta P_i + \epsilon_i \qquad (4.1)$$

dove ϵ_i denota un residuo incorrelato con P_i, la varianza dello stimatore di Horvitz-Thompson è proporzionale a

$$E\left(\sum_i \frac{\epsilon_i}{nP_i}\right)^2 \qquad (4.2)$$

mentre l'errore quadratico medio dello stimatore basato sul quoziente è proporzionale a

$$E\left(\frac{\Sigma\epsilon_i}{\Sigma P_i}\right)^2, \qquad (4.3)$$

ed è intuitivo che le varianze dei possibili campioni sono più omogenee tra loro perlomeno quando i residui hanno segni tra loro diversi. Inoltre, la prossimità di $v_s(r)$ a $V(r)$ è tanto più stretta quanto meno fluttuano le probabilità di inclusione (v. anche Kish, 1965: 6.6). Tuttavia, non sempre la stratificazione per dimensione dà varianze più stabili di quella finalizzata allo svolgimento dell'indagine sulle forze di lavoro.

La stabilità dello stimatore dell'errore quadratico medio applicato allo stimatore del totale (formula 2.60) tende ad essere considerevolmente superiore a quella degli stimatori specificatamente proposti per le varie tecniche e a quello asintotico per stime lineari proposto da Hartley e Rao (1962), sia con la stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro, sia — ma in misura quasi sempre meno elevata — per la stratificazione per dimensione dei comuni.

In appresso, si riportano sintetiche considerazioni sulla stabilità degli stimatori della varianza associati alle singole tecniche di formazione dei campioni.

Tecnica 2. Lo stimatore proposto da Hartley e Rao (1962) per la varianza di stime lineari è generalmente più stabile della tecnica con reimmissione. È in Toscana, dove il numero di comuni per strato è mediamente basso, e soprattutto con la stratificazione per dimensione, che le stime della varianza si discostano meno dal loro valor medio. Per $n = 3$, la stabilità delle stime è decisamente superiore a quella ottenibile con il campionamento con reimmissione, e, sempre $n = 3$, la stratificazione per dimensione dà risultati generalmente migliori.

Spesso la stima basata sul quoziente con campioni di comuni selezionati sistematicamente da strati internamente ordinati dà stime meno stabili del campionamento con reimmissione.

Tecnica 3. La stabilità delle stime della varianza derivanti dall'applicazione della formula asintotica di Hartley e Rao (1962) supera quella del campionamento con reimmissione. Il campionamento sistematico da liste casuali è, invece, praticamente equivalente a quello bernoulliano quando si applica uno stimatore basato sul quoziente.

Tecnica 4. La selezione senza reimmissione con la procedura codificata da Brewer (1963) dà stime nel complesso più stabili di quelle ottenute con il procedimento proposto da Rao, Hartley e Cochran (1962) e, naturalmente, del campionamento con reimmissione, ma meno stabile di quasi ogni altra tecnica. Le probabilità di inclusione proposte da Brewer non si applicano con successo a stime basate sul quoziente tra somme.

Tecnica 5. La stabilità delle varianze dei campioni ottenibili con la tecnica del raggruppamento di Durbin (1967) non è stata calcolata per le ragioni esposte nel par. 3.

Tecnica 6. Il metodo iterativo di Fellegi (1963) dà stime considerevolmente stabili, soprattutto con campioni formati entro strati definiti in base alla dimensione demografica dei comuni.

La dimensione dello strato non sembra legata all'efficienza del metodo; la convergenza non dipende, cioè, se non nella tempestività, dall'essere lo strato di dimensioni contenute. Si osserva, invece, che le differenze in stabilità rispetto alla tecnica con reimmissione di Brewer-Murthy, l'unica sistematicamente più stabile di quella di cui si tratta, si riducono negli strati più omogenei per dimensione, e quando la frazione di campionamento è più elevata.

Tecnica 7. Per campioni di coppie di comuni ottenuti da substrati formati con una tecnica analoga a quella ideata da Rao, Hartley e Cochran (1962), la stabilità relativa delle stime della varianza è piuttosto simile in valore all'efficienza degli stimatori del totale, ossia di poco superiore a quella del campionamento con reimmissione e inferiore praticamente a tutte le tecniche di selezione in blocco delle unità. Talvolta si ottengono varianze anche meno stabili di quelle del campionamento con reimmissione. Questo risultato contrasta nettamente con le conclusioni di Rao e Bayless (1969) che vedono lo stimatore di Rao-Hartley-Cochran costantemente, e talvolta in modo appariscente, in posizione favorevole per quanto riguarda la stabilità della varianza rispetto a tutti quelli considerati (tra gli altri, quello

di Murthy (1957), di Raj (1956), di Fellegi (1963) e di quello con reinserimento).

Elaborando campioni di numerosità 3 e 4, Bayless e Rao (1970) trovano ulteriore conferma della stabilità della tecnica di cui si tratta, esclusi alcuni casi caratterizzati da forte variabilità nelle probabilità di selezione e da elevata frazione di campionamento. Per $n = 3$ anche le varianze ricavate dai nostri sub-strati sono considerevolmente stabili e l'efficienza varia, mediamente sulle 10 variabili considerate, dal 27% in più rispetto al campionamento con reimmissione con la prima stratificazione in Piemonte al 78% in più con la stratificazione alternativa sempre in Piemonte.

Tecnica 8. Per la tecnica di estrazione senza reimmissione di Yates e Grundy (1953) si è calcolata la stabilità del solo stimatore proposto da Murthy (1957) ([4]). Questo stimatore è decisamente il più stabile tra quelli considerati nelle elaborazioni svolte, sia con la stratificazione che raggruppa i comuni più omogenei per dimensione demografica dentro la provincia, sia con quella pertinente all'indagine sulle forze di lavoro. La misura è circa due volte e mezza quella del campionamento con reimmissione per $n = 2$.

Applicato a campioni selezionati senza reimmissione con probabilità determinate dalla formula di Brewer (1963), Rao e Singh (1973) confermano che lo stimatore di Murthy ha varianza molto stabile.

Tecnica 9. Nonostante che la varianza dello stimatore applicato alla tecnica di rifiuto di campioni non distinti sia spesso elevata, le stime della varianza sono considerevolmente stabili. Calcolata con la stratificazione per dimensione dei comuni, la stabilità della varianza è inferiore solo a quella inerente alla tecnica dei sub-strati casuali di Rao-Hartley-Cochran in Toscana e a quella di selezione in blocco con lo stimatore di Murthy nelle tre regioni esaminate.

Tecnica 10. È una tecnica che genera stime della varianza piuttosto instabili. Talvolta è perfino meno stabile della tecnica di selezione con reimmissione.

Funziona meglio quando la probabilità di inclusione variano poco entro gli strati.

Tecnica 11. La stabilità del campionamento con probabilità costanti è decisamente inferiore a quella del campionamento con pro-

babilità variabili. Solo con la stratificazione per dimensione e per strati di numerosità sufficientemente grande, ovvero quando la variabilità delle probabilità di estrazione nel campionamento PPS è contenuta, si ottengono stime con un grado di stabilità paragonabile alle altre considerate (il massimo è il 74% circa in Calabria in rapporto al campionamento con reimmissione).

5. APPLICABILITÀ DELLE TECNICHE ADOTTATE

Nella Tavola 7 in Appendice si riportano alcune informazioni di larga massima sui problemi tecnici relativi alla formazione di campioni di due comuni per strato e alla stima delle statistiche essenziali.

Si nota che:

a) quantunque la misura del tempo di esecuzione dei programmi per il calcolo automatico sia approssimata, i tempi di esecuzione dei calcoli risultano sistematicamente più elevati per il Piemonte, a causa del maggior numero di comuni che a questa regione appartengono. I campioni possibili con la tecnica di selezione con reimmissione sono oltre 30 mila, con una media per strato di 604 (su 48 strati) quando si adotta la stratificazione per dimensione dei comuni e di 361 (su 46 strati) con la stratificazione derivata dall'indagine sulle forze di lavoro. Anche se per l'Istat, che dispone di un autonomo Centro di Calcolo, il tempo di occupazione del *computer* può non essere sentito come un problema, qualora si adotti uno schema di selezione del campione che preveda di individuare $\binom{N}{n}$ campioni distinti di n unità, è utile anche il contenimento della variabilità numerica dei comuni per strato. Se N non variasse da strato a strato, il numero dei possibili campioni sarebbe, infatti, in Piemonte, attorno a 14.500;

b) la tecnica n. 6, per la quale si determina iterativamente la probabilità di inclusione nel campione di ogni unità, è quella che assorbe in assoluto la maggiore quantità di tempo-macchina, seguita dalle procedure n. 5 e 7, per le quali è prevista la suddivisione di ogni strato in sub-strati prima della selezione. Anche in quest'ultimo caso è la numerosità dei possibili campioni che determina i tempi di esecuzione;

c) la tecnica n. 2, anche se richiede un certo tempo per la generazione degli N possibili campioni, è quella che, tra tutte le tecniche esaminate, individua il campione da analizzare nel minor tempo, richiedendo anche un limitatissimo spazio di memoria;

d) le probabilità generate sono positive per tutte le tecniche, ad eccezione delle tecniche n. 5 e 10.

La prima delle due genera probabilità negative in corrispondenza della stratificazione derivante dal piano di campionamento dell'indagine sulle forze di lavoro nella selezione della seconda unità dello stesso gruppo della prima. Le probabilità negative scompaiono se si stratifica per dimensione.

Anche la tecnica n. 10 beneficia della stratificazione per dimensione, tuttavia le probabilità negative sono numerose e sparse su molti strati se non si adotta la correzione già menzionata. Con l'altra stratificazione non sono a rigore confrontabili i risultati ottenuti con queste due tecniche e con le altre in 35 strati su 46 in Piemonte e in 10 su 22 in Toscana;

e) per la tecnica n. 10 può porsi il problema della disponibilità di memoria, anche disponendo di un calcolatore di grandi dimensioni. Mentre, infatti, per ogni altra tecnica considerata la selezione è sequenziale — ossia si può immaginare di estrarre una unità dopo l'altra, e in tal caso la memoria è dimensionabile in funzione della numerosità complessiva dei comuni da sottoporre a campionamento — la tecnica di selezione di campioni interi, così com'è proposta, richiede un dimensionamento in funzione della stratificazione. Se si creano strati di grandi dimensioni, può essere necessario modificare il programma di selezione;

f) le tecniche di selezione senza reimmissione (n. 8) e di rifiuto di campioni non distinti (n. 9) congiuntamente proposte da Yates e Grundy (1954) non sono «esatte» perché in genere $\pi_i \neq nP_i$. La variabilità dell'espressione $d_i = (\pi_i - nP_i) / nP_i$ è stata calcolata per $n = 2$ per la tecnica n. 8:

$$V(d') = \frac{1}{4N} \sum_{i}^{N} \left[\sum_{k \neq i}^{N} \frac{P_k}{1 - P_k} - 1 \right]^2 \qquad (5.1)$$

e per la tecnica n. 9:

$$V(d'') = \frac{1}{N} \sum_{i}^{N} \left[\frac{1 - P_i}{1 - \sum_k P_k^2} - 1 \right]^2 . \qquad (5.2)$$

I risultati per la stratificazione basata sul campione in uso per l'indagine sulle forze di lavoro danno 1000 V(d') = 2,016 per la Toscana e 1000 V(d') = 1,426 per il Piemonte; con la stratificazione per dimensione la variabilità si riduce considerevolmente: 1000 V(d') = 0,399 per la Toscana e 1000 V(d') = 0,055 per il Piemonte.

Per la tecnica n. 9, i risultati con la prima stratificazione sono: 1000 V(d'') = 11,889 per la Toscana e 1000 V(d'') = 7,687 per il Piemonte. Anche in questo caso la variabilità scende se si stratifica tenendo conto della dimensione e si ottiene 1000 V(d'') = 2,147 per la Toscana e 1000 V(d'') = 0,384 per il Piemonte.

Sia per la tecnica n. 8, sia per la n. 9, le differenze tendono a scomparire con l'aumentare della numerosità negli strati. Yates e Grundy (1953) sostengono che si può introdurre la probabilità $2P_i$ invece della corretta π_i, considerando trascurabile la distorsione introdotta da questa scelta, anche adottando lo stimatore di Horvitz-Thompson.

6. CONSIDERAZIONI CONCLUSIVE

I commenti alle elaborazioni condotte sull'archivio di dati comunali permettono di trarre alcune considerazioni propositive per la messa in opera di un programma di campionamento «generalizzato» che permetta di scegliere la stratificazione, la numerosità campionaria, la tecnica di formazione del campione e lo stimatore delle statistiche che interessano.

Per quanto attiene alla stratificazione sono emerse varie indicazioni favorevoli a una suddivisione delle unità in strati caratterizzati da omogeneità nella dimensione dei comuni.

Analizzando i dati presentati, abbiamo maturato la convinzione che nel progettare un'indagine che si proponga di analizzare diverse variabili, la maggior parte delle quali ha una distribuzione ignota, conviene adottare regole di campionamento semplici purché diano stime prossime all'ottimalità (una discussione sulla ricerca di *proxi-*

ma, invece di *optima*, nella progettazione di campioni si trova in Kish, 1976).

Per quanto concerne la stratificazione, se ci è permesso esprimerci con una locuzione figurata, si può puntare a costruire blocchi di unità di forma regolare e con pochi spigoli. È un approccio affatto diverso, *inter alia*, della stratificazione fondata su tecniche di raggruppamento multivariate, le quali generano partizioni delle unità di numerosità non controllata, talvolta molto variabile, con tanti gruppi contenenti solo una o due unità. Tra l'altro, se si presta fede alle analisi di Biggeri *et al.* (1977) e di Zani e Sicuri (1977 a; 1977 b) i risultati della stratificazione multivariata sulle stime sono modesti.

Con stratificazioni «ottimali» nel significato attribuitogli da Neyman si formano classi di unità miranti ad ottimizzare la varianza di una o più statistiche, ma i risultati valutati sull'insieme delle variabili sono verosimilmente non superiori a quelli della estrazione con probabilità proporzionali alla dimensione, come si congettura nel par. 3.

Con l'attenzione principalmente rivolta all'efficienza delle stime, e con in mente lo stimatore di Horvitz-Thompson (formula 2.28), si possono studiare stratificazioni più elaborate di quella adottata nello svolgere le elaborazioni presentate. Per esempio, suddivisioni delle unità in H strati di N_h unità ($h = 1,..., H$) che minimizzino la varianza dello stimatore, contenendo, allo stesso tempo, la variabilità numerica dei comuni in strati diversi e la variabilità nella dimensione demografica degli strati. Si ricorda che nel formare gli strati per dimensione si è tenuto conto solo di questo ultimo criterio.

Per quanto riguarda la numerosità campionaria, non abbiamo trovato nella letteratura consultata (tra gli altri: Shapiro e Olsen, 1979; Tadros *et al.*, 1982; Gosh, 1983), né nelle analisi prodotte, alcuna indicazione contraria a selezionare due comuni per strato ogni qualvolta se ne presenti l'opportunità.

Le valutazioni delle varie tecniche sono qui riprese in chiave propositiva.

Tecnica 1. La selezione con reimmissione è presa come base per il confronto della efficienza degli stimatori e della stabilità delle varianze delle altre tecniche. Quantunque sia una tecnica efficiente e la varianza non sia mai negativa, non è una tecnica appropriata quando si vogliano campioni di unità distinte.

Tecnica 2. Il campionamento sistematico con ordinamento delle unità dentro gli strati (Madow, 1949) è una tecnica già attiva su un calcolatore dell'Istat, e facilmente generalizzabile per qualsiasi numerosità di campionamento. È, inoltre, di rapida esecuzione e intuitiva per ogni utente di un eventuale programma di selezione automatica dall'archivio di comuni.

Con lo stimatore (asintotico) proposto da Hartley e Rao (1962) è anche altamente efficiente e la varianza è stabile in prove ripetute, soprattutto se si adotta una stratificazione basata sulla dimensione delle unità. È una delle tecniche che si possono raccomandare per essere utilizzate in qualsiasi realtà operativa.

Tecnica 3. È ancora una tecnica di campionamento sistematico, ma le unità sono disposte casualmente nella lista. Per stimare la varianza si è fatto ricorso a uno stimatore asintotico che ha dato i migliori risultati quando le unità sono poste in sequenza secondo la dimensione. Se, dunque, si adotta il campionamento sistematico, a questa tecnica, è preferibile la tecnica n. 2.

Tecnica 4. La tecnica di selezione «in blocco» con procedura suggerita da Brewer (1963) per $n = 2$ e sempre da Brewer (1975) per n qualsiasi è altamente efficiente anche se le stime non sono molto stabili. Come efficienza, le stime si collocano al livello di quelle ricavate con la tecnica n. 2 ma la stabilità è superiore solo al campionamento bernoulliano e a quello effettuato da sub-strati casuali (tecnica n. 7).

Tecnica 5. La tecnica di formazione previa di gruppi di comuni suggerita da Durbin (1967) è dispendiosa, difficilmente applicabile (genera talvolta probabilità negative), anche se sufficientemente affidabile nella stima. Non se ne consiglia l'adozione.

Tecnica 6. La determinazione delle probabilità di inclusione con la tecnica di ottimizzazione proposta da Fellegi (1963), pur implicando lo sfruttamento di risorse informatiche relativamente cospicue, dà stime considerevolmente efficienti e stabili nella efficienza in prove ripetute. In alcuni strati, però, la procedura iterativa adottata (Brewer e Hanif, 1983; app. A) non converge e l'algoritmo non è stato provato per numerosità superiori a 2.

Siccome la tecnica è stata proposta per risolvere specificatamente, anche se non solo, il problema della fluttuazione della probabilità di inclusione in prove successive con campioni ruotati, non

troviamo argomenti per consigliarla nella formazione di campioni generalizzati da utilizzare per una sola occasione di indagine.

Tecnica 7. La procedura dei sub-strati casuali di Rao, Hartley e Cochran (1962) è una tecnica che dà stime moderatamente efficienti, di varianza abbastanza stabile, ed è facilmente applicabile per qualsiasi numerosità campionaria. I sub-strati casuali si formano con criteri di rappresentatività dello strato e la suddivisione casuale delle unità è tanto più agevole quanto maggiore è la numerosità dello strato. I benefici in efficienza si stemperano al divergere della numerosità campionaria e l'unico vantaggio rispetto al campionamento con reimmissione sta nella generazione di campioni sempre distinti.

D'altro canto, rispetto alle tecniche di selezione sistematica e in blocco, abbiamo ottenuto valori di efficienza generalmente inferiori (*contra*: Rao e Bayless, 1969 e Bayless e Rao, 1970).

Dalle elaborazioni svolte emergono indicazioni favorevoli all'abbinamento della tecnica di cui si tratta con la stratificazione per dimensione. Per formare campioni «tranquilli», conviene sperimentare ulteriormente le condizioni che ne migliorano la resa (altre variabili, altre regioni, varie stratificazioni).

Tecnica 8. Anche questa tecnica proposta da Yates e Grundy (1953) prevede una selezione in blocco delle unità. Con gli stimatori proposti da Raj (1956) e, più ancora, con quello proposto da Murthy (1957) si ottengono stime altamente efficienti, stabili e di facile applicazione per $n = 2$. Per numerosità superiori, le formule della varianza richiedono un impegno di programmazione informatica evitato in questa istanza. Inoltre, la probabilità di selezione di una unità non è esattamente uguale a nP_i, ma fluttua, anche se in misura contenuta. La variabilità nella probabilità di inclusione dei comuni si ripercuote sull'intero disegno di campionamento se il campione è selezionato su più stadi.

Con il campionamento in blocco, per campioni di numerosità 2 per strato, ci si può orientare indifferentemente sugli stimatori proposti da Horvitz e Thompson (1952) e da Murthy (1957) con probabilità di selezione date, rispettivamente, da Brewer (1963) e Yates e Grundy (1953). Per campioni di numerosità superiore, *rebus sic stantibus*, si consiglia di seguire la via dello stimatore convenzionale di Horvitz-Thompson con le probabilità di selezione indicate da Brewer (1975).

Tecnica 9. La tecnica di rifiuto dei campioni non distinti di Yates e Grundy (1953) è piuttosto inefficiente, anche se le stime della varianza sono stabili. È una tecnica che, se si considera anche l'artificiosità della procedura di selezione su calcolatore, non è preferibile ad altre trattate.

Tecnica 10. La tecnica di selezione di campioni interi proposta da Herzel (1984) genera talvolta probabilità negative. Lo stesso Herzel suggerisce una procedura che pone rimedio a questo inconveniente; dal punto di vista informatico questa procedura è a tutt'oggi un problema aperto.

La versione adottata nelle elaborazioni svolte è sconsigliabile per formare campioni. L'interesse primario per rendere concorrenziale la tecnica va posto sulla realizzazione di una procedura informatica che risolva automaticamente il problema delle probabilità «selvagge».

Tecnica 11. Il campionamento casuale semplice è decisamente non comparabile con le tecniche esposte per quanto riguarda l'efficienza degli stimatori e delle loro varianze. Per campioni a più stadi autoponderanti la sua applicazione è, inoltre, sconsigliata perché fa perdere il controllo sulla numerosità campionaria ai successivi stadi.

Riepilogando ulteriormente, al fine di formare campioni «generalizzati» per indagini che si svolgono in una sola occasione, si consiglia l'impiego della tecnica sistematica (n. 2) per campioni di qualsiasi numerosità e la tecnica senza reimmissione (n. 8) con lo stimatore di Murthy (1957) per campioni non superiori a 3.

Per analisi più approfondite si indica di studiare la relazione esistente tra numerosità, misura di grandezza dello strato, coefficienti di variazioni di queste ed altre variabili e della probabilità di inclusione, coefficiente di correlazione tra variabili e dimensione delle unità, da una parte, e varie statistiche di misura della affidabilità delle stime, della efficacia del campionamento (probabilità sempre positive, ecc.) e delle risorse richieste per il campionamento dall'altra.

Un altro percorso, che nel lavoro presentato ha dato risultati solo parziali, ma che è caldeggiato da molti studiosi di tecniche campionarie, è quello dello stimatore basato sul quoziente. Le prove sono state svolte con riferimento allo stimatore «convenzionale»; Tin (1965) fornisce le formule della distorsione, dell'efficienza e della conve-

nienza computazionale di vari stimatori alternativi basati sul quoziente (quelli proposti da: Quenouille, 1956; Beale, 1962; Hartley e Ross, 1954; dallo stesso Tin e uno stimatore modificato basato sul quoziente); Rao e Rao (1971), Brewer (1979) e Royall e Cumberland (1981), suggeriscono stimatori che presuppongono l'adattamento di modelli (di superpopolazione) nel calcolo della varianza. Si tratterebbe, comunque, di analisi di completamento, volte ad affinare le prospettive, più che a sostituirsi a quanto fino a questo punto descritto.

TAVOLE RIASSUNTIVE DELLE ELABORAZIONI SVOLTE

LEGENDA DEI NUMERI CHE CONTRADDISTINGUONO
LA TECNICA DI CAMPIONAMENTO NELLE TESTATE
DELLE TAVOLE CHE SEGUONO

1: Selezione con reimmissione delle unità estratte
2: Campionamento sistematico con stratificazione implicita dei comuni e stimatore proposto da Hartley (1966)
3: Campionamento sistematico con casualizzazione delle posizioni dei comuni nella lista
4: Selezione senza reimmissione di Brewer (1963)
5: Metodo del raggruppamento di Durbin (1967)
6: Metodo iterativo di Fellegi (1963)
7: Procedura degli strati casuali di Rao, Hartley e Cochran (1962)
8: Estrazione senza reimmissione di Yates e Grundy (1953)
9: Tecnica di rifiuto dei campioni non distinti di Yates e Grundy (1953)
10: Tecnica di determinazione della probabilità di estrazione di campioni interi di Herzel (1984)
11: Campionamento casuale semplice senza reimmissione.

Tavola 1 - Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, con stratificazione dei comuni derivata da quella in uso per l'indagine sulle forze di lavoro, per tecnica di campionamento dei comuni (c), stimatore adottato variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	7	8 (a)	8 (b)	9	10	11
TOSCANA										
1. Popolazione disoccupata										
St. A	100	120,2	114,4	118,6	111,5	115,3	119,0	114,3	118,9	21,5
St. B	95,4	119,7	112,3	105,8	102,9	115,5	=	110,0	115,9	26,2
2. Femmine disoccupate										
St. A	100	114,9	114,2	117,8	111,4	115,3	118,9	117,7	117,7	28,7
St. B	97,4	117,3	116,3	112,0	104,4	«119,8	=	112,4	119,9	34,9
3. Popolazione occupata										
St. A	100	141,2	111,6	114,5	109,8	112,5	114,9	58,8	112,1	0,1
St. B	98,5	136,0	112,7	26,3	108,6	44,1	=	100,0	110,5	1,2
4. Femmine occupate										
St. A	100	136,1	110,6	112,8	109,1	111,5	113,5	86,6	108,8	4,2
St. B	99,0	129,1	111,9	75,8	108,2	94,6	=	110,3	108,7	4,7
5. Popolazione in condizione profess.										
St. A	100	138,4	111,5	114,1	109,7	112,4	114,8	57,2	111,2	0,1
St. B	99,6	135,8	114,1	25,5	109,5	43,4	=	100,2	111,1	1,2
6. Femmine in cond. profess.										
St. A	100	133,9	110,7	112,9	109,2	111,7	113,7	86,0	108,6	3,9
St. B	99,5	127,6	112,9	74,4	108,7	94,2	=	110,8	109,4	4,6
7. Residenti con scuola obbligo										
St. A	100	116,8	111,6	114,5	109,1	112,2	114,4	53,9	110,6	1,4
St. B	92,7	107,1	103,6	42,6	103,6	63,7	=	105,6	101,8	1,6
8. Femmine con scuola obbligo										
St. A	100	116,4	110,8	113,2	108,8	111,6	113,5	63,0	110,9	1,9
St. B	96,0	114,4	107,1	53,2	105,9	74,5	=	108,7	106,3	2,3
9. Popolazione con laurea										
St. A	100	147,5	111,2	113,9	107,8	110,8	112,4	104,2	112,9	24,9
St. B	81,7	138,1	89,1	99,5	84,6	97,5	=	90,6	93,7	28,0
10. Femmine laureate										
St. A	100	147,0	111,2	114,1	107,9	110,9	112,6	103,7	112,0	24,0
St. B	82,6	132,8	90,1	103,1	86,0	99,9	=	92,0	93,7	28,0

(a) Stimatore proposto da Raj (1956). - (b) Stimatore della varianza proposto da Murthy (1957) per lo stimatore del totale proposto da Ray (1956). - (c) La varianza inerente alle tecniche n. 5 e n. 6 non è calcolabile per alcuna regione.

Tavola 1 segue - Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, con stratificazione dei comuni derivata da quella in uso per l'indagine sulle forze di lavoro, per tecnica di campionamento dei comuni (c), stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	7	8 (a)	8 (b)	9	10	11
PIEMONTE										
1. Popolazione disoccupata										
St. A	100	102,2	106,3	107,4	103,5	107,1	107,9	98,7	100,1	9,2
B	91,2	91,0	96,2	108,3	93,8	104,2	=	96,9	92,0	9,8
2. Femmine disoccupate										
St. A	100	103,8	105,9	106,9	103,4	106,8	107,5	100,1	100,5	11,3
B	93,8	92,7	98,8	111,0	95,6	106,5	=	99,3	93,1	12,1
3. Popolazione occupata										
St. A	100	113,7	104,8	105,5	103,5	106,4	107,0	43,9	97,8	0,2
B	110,2	123,6	117,2	26,0	115,8	69,7	=	104,3	96,0	0,3
4. Femmine occupate										
St. A	100	111,5	104,9	105,6	103,4	106,3	107,0	70,5	98,9	0,8
St. B	108,7	123,7	115,4	65,4	114,3	106,8	=	112,7	96,7	0,9
5. Popolazione in condizione profess.										
St. A	100	114,7	104,7	105,4	103,4	106,3	107,0	41,2	95,6	0,2
St. B	109,8	122,0	116,6	26,7	114,9	72,7	=	106,0	93,9	0,3
6. Femmine in cond. profess.										
St. A	100	113,8	104,9	105,7	103,5	106,4	107,0	67,7	97,4	0,8
St. B	107,5	123,3	114,0	67,4	112,3	109,1	=	112,6	94,7	0,8
7. Residenti con scuola obbligo										
St. A	100	119,1	105,6	106,9	103,6	106,8	107,5	61,3	82,7	1,0
St. B	100,0	123,3	104,6	103,5	103,7	128,4	=	112,2	78,9	1,0
8. Femmine con scuola obbligo										
St. A	100	121,6	105,7	107,1	103,6	106,8	107,5	66,7	96,1	1,3
St. B	100,1	127,2	104,7	117,8	103,7	128,0	=	112,2	81,2	1,3
9. Popolazione con laurea										
St. A	100	109,6	109,8	114,4	104,6	109,7	111,4	95,8	106,7	7,7
St. B	79,9	75,5	84,4	104,8	86,3	95,0	=	89,1	84,8	8,3
10. Femmine laureate										
St. A	100	107,5	108,9	113,1	104,1	109,0	110,5	97,1	104,4	104,4
St. B	82,8	79,7	86,7	108,2	88,4	97,5	=	91,6	83,8	83,8

(a), (b), (c) vedi nota a pag. 127.

Tavola 1 segue - **Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, con stratificazione dei comuni derivata da quella in uso per l'indagine sulle forze di lavoro, per tecnica di campionamento dei comuni** (c), **stimatore adottato, variabile di riferimento e regione** *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8 (a)	8 (b)	9	10	11
CALABRIA											
1. Popolazione disoccupata											
St. A	100	108,8	109,4	111,1	105,5	108,6	111,1	113,0	116,5	110,0	37,2
St. B	101,3	102,1	113,5	108,3	108,4	108,0	113,7	110,0	113,5	40,4	
2. Femmine disoccupate											
St. A	100	104,6	108,6	109,9	107,8	108,4	110,7	112,3	114,8	108,9	50,7
St. B	106,5	115,1	119,2	118,8	115,5	111,3	122,0	115,5	116,5	55,2	
3. Popolazione occupata											
St. A	100	113,6	109,2	111,1	88,2	107,5	110,4	112,0	82,5	108,8	2,6
St. B	99,8	103,2	110,5	65,7	88,0	106,9	92,7	107,6	108,0	2,8	
4. Femmine occupate											
St. A	100	114,5	110,3	112,9	103,1	107,4	110,9	112,6	110,2	111,2	13,9
St. B	91,0	93,2	99,8	91,3	97,2	101,29	9,0	99,7	15,0		
5. Popolazione in condizione profess.											
St. A	100	116,3	109,1	111,1	69,7	107,5	110,4	111,9	81,5	107,9	1,9
St. B	99,4	111,7	109,5	47,0	72,8	104,5	74,7	102,3	104,9	2,0	
6. Femmine in cond. profes.											
St. A	100	111,6	110,4	113,2	97,4	107,7	111,1	112,9	112,8	110,7	12,5
St. B	90,9	96,2	99,6	83,5	89,4	96,8	96,6	98,1	98,4	13,5	
7. Residenti con scuola obbligo											
St. A	100	113,0	111,2	114,8	86,7	107,5	111,6	113,7	79,2	114,9	2,2
St. B	91,2	103,9	101,6	57,9	82,1	103,7	81,4	100,6	103,3	2,4	
8. Femmine con scuola obbligo											
St. A	100	116,8	110,8	114,2	89,1	107,6	111,4	113,5	87,7	114,5	2,9
St. B	93,6	105,6	104,8	64,9	86,4	103,3	87,5	102,8	107,3	3,2	
9. Popolazione con laurea											
St. A	100	108,9	112,8	116,0	124,1	107,1	112,4	114,4	103,0	116,3	16,0
St. B	77,3	65,5	84,8	98,3	88,5	82,1	94,9	86,2	90,6	17,1	
10. Femmine laureate											
St. A	100	111,9	112,0	114,7	122,2	107,1	112,0	113,9	104,4	114,6	18,8
St. B	81,3	71,7	89,8	103,5	93,0	85,4	99,8	90,2	95,6	20,1	

(a), (b) e (c) vedi nota a pag. 127.

Il valore assoluto della varianza delle stime ottenute con la tecnica 1 è:

	M+F	F
	TOSCANA	
Popolazione disoccupata	979.191	404.013
Popolazione occupata	37.293.161	18.793.810
Popol. in cond. professionali	37.174.813	19.795.205
Residenti con scuola obbligo	13.424.956	3.544.278
Residenti con laurea	2.460.602	398.854
	PIEMONTE	
Popolazione disoccupata	1.666.911	595.546
Popolazione occupata	41.567.737	18.670.542
Popol. in cond. professionali	42.629.672	19.577.289
Residenti con scuola obbligo	54.854.947	15.959.187
Residenti con laurea	2.325.125	404.608
	CALABRIA	
Popolazione disoccupata	28.936.158,5	12.411.518,6
Popolazione occupata	44.772.802,6	20.757.384,6
Popol. in cond. professionali	42.042.268,2	27.342.647,7
Residenti con scuola obbligo	16.640.541,7	3.862.093,1
Residenti con laurea	2.716.555,7	593.516,8

Tavola 2 - Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, con stratificazione per dimensione demografica, per tecnica di campionamento (c), stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8 (a)	8 (b)	9	10	11
TOSCANA											
1. Popolazione disoccupata											
St. A	100	112,3	115,6	119,4	119,3	120,1	116,5	121,1	119,4	119,2	45,8
St. B	104,3	128,8	(d)	124,7	124,7	128,2	128,2	=	127,7	126,4	57,4
2. Femmine disoccupate											
St. A	100	103,7	115,0	118,3	117,6	120,0	116,1	120,6	121,2	117,7	55,1
St. B	104,1	111,3	(d)	126,4	123,9	123,0	129,3	=	126,3	126,0	68,2
3. Popolazione occupata											
St. A	100	133,7	111,9	114,6	90,6	112,8	112,3	114,8	83,7	110,9	3,1
St. B	101,3	130,9	(d)	30,3	93,6	114,9	35,2	=	104,5	112,1	3,6
4. Femmine occupate											
St. A	100	127,8	110,9	113,1	110,3	111,6	111,2	113,3	101,1	108,9	11,3
St. B	100,0	125,6	(d)	74,6	108,9	109,5	80,1	=	111,1	109,2	13,0
5. Popolazione in condizione profess.											
St. A	100	135,0	111,9	114,7	90,0	112,9	112,4	115,0	82,6	110,8	3,0
St. B	101,9	133,1	(d)	29,3	93,4	115,5	34,2	=	104,7	112,7	3,4
6. Femmine in cond. profes.											
St. A	100	130,4	111,1	113,4	110,2	111,9	111,4	113,6	101,1	108,9	10,9
St. B	100,7	130,1	(d)	73,4	109,7	111,3	79,1	=	112,0	110,0	12,6
7. Residenti con scuola obbligo											
St. A	100	125,0	112,9	116,1	109,9	114,6	113,4	116,6	85,3	115,4	4,5
St. B	102,8	127,9	(d)	41,7	110,3	117,1	46,8	=	114,5	118,6	5,3
8. Femmine con scuola obbligo											
St. A	100	116,0	111,7	114,2	107,1	112,9	112,2	114,7	91,8	113,7	6,3
St. B	102,6	120,3	(d)	50,1	108,6	116,6	55,8	=	113,2	116,6	7,3
9. Popolazione con laurea											
St. A	100	104,6	113,6	116,5	118,6	114,1	113,4	116,0	112,6	116,2	56,9
St. B	93,9	95,8	(d)	116,4	108,5	98,8	116,1	=	109,3	107,8	71,3
10. Femmine laureate											
St. A	100	111,8	113,6	116,6	119,2	114,2	113,4	116,1	112,3	116,5	56,6
St. B	93,9	98,2	(d)	117,6	108,7	98,8	117,0	=	109,4	107,9	70,9

(a) Stimatore proposto da Raj (1956); (b) Stimatore della varianza proposto da Murthy (1957) per lo stimatore del totale proposto da Raj (1956); (c) la varianza inerente alla tecnica n. 5 non è calcolabile per alcuna regione; (d) alcuni strati hanno numerosità 3 e ciò rende impraticabile l'applicazione della formula della varianza asintotica di Hartley e Rao (1962).

Tavola 2 segue - **Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, con stratificazione per dimensione demografica, per tecnica di campionamento** (c), **stimatore adottato, variabile di riferimento e regione** *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO										
	1	2	3	4	6	7	8 (a)	8 (b)	9	10	11
PIEMONTE											
1. Popolazione disoccupata											
St. A	100	111,5	108,6	110,3	109,0	110,1	108,8	110,5	110,6	110,2	86,7
St. B	101,5	115,6	112,4	107,3	111,1	117,4	109,0	=	112,5	111,6	103,2
2. Femmine disoccupate											
St. A	100	114,6	108,6	110,3	109,1	110,1	108,7	110,6	110,7	110,3	87,0
St. B	101,0	117,6	111,7	108,7	110,5	116,6	110,3	=	111,8	111,6	103,0
3. Popolazione occupata											
St. A	100	106,6	106,9	108,0	79,6	107,6	107,0	108,1	93,9	108,0	11,6
St. B	100,2	97,8	108,1	23,7	81,1	106,8	26,9	=	91,9	107,8	13,6
4. Femmine occupate											
St. A	100	109,8	107,2	108,3	101,9	108,1	107,3	108,5	102,1	108,4	30,5
St. B	99,5	107,4	107,7	57,2	100,9	105,7	61,3	=	103,8	108,0	35,9
5. Popolazione in condizione profess.											
St. A	100	105,6	106,9	107,9	78,0	107,5	106,9	108,0	93,3	107,9	11,1
St. B	100,3	96,5	108,2	22,5	79,7	107,0	25,6	=	90,9	107,8	13,0
6. Femmine in cond. profes.											
St. A	100	107,6	107,2	108,2	101,1	108,1	107,3	108,5	102,0	108,3	29,7
St. B	99,5	104,8	107,6	55,4	100,2	106,0	59,5	=	103,3	107,9	34,9
7. Residenti con scuola obbligo											
St. A	100	103,4	107,6	108,9	92,3	108,4	107,6	109,0	103,4	108,8	27,0
St. B	100,7	100,4	109,7	41,6	94,3	109,1	46,0	=	101,3	109,2	32,0
8. Femmine con scuola obbligo											
St. A	100	100,6	107,3	108,4	98,9	107,9	107,4	108,5	101,9	108,2	32,1
St. B	100,4	99,3	108,8	51,1	99,0	110,7	55,3	=	103,4	108,3	39,1
9. Popolazione con laurea											
St. A	100	108,3	110,8	112,8	113,5	111,2	110,7	112,6	110,9	112,9	83,7
St. B	97,1	104,8	108,8	113,9	108,9	110,0	114,5	=	110,1	108,8	104,5
10. Femmine laureate											
St. A	100	109,6	110,4	112,5	113,5	111,1	110,3	112,3	110,5	112,5	84,1
St. B	98,1	107,4	109,9	114,0	108,9	110,6	114,7	=	111,0	109,7	104,8

(a), (b), (c) e (d) vedi nota a pag. 131.

Tavola 2 segue - **Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, con stratificazione per dimensione demografica, per tecnica di campionamento** (c), **stimatore adottato, variabile di riferimento e regione** *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8 (a)	8 (b)	9	10	11
CALABRIA											
1. Popolazione disoccupata											
St. A	100	107,5	111,6	114,1	113,9	113,0	111,6	113,9	113,9	114,1	94,8
St. B	99,0	106,7	112,7	121,5	112,7	113,4	121,6		114,0	112,7	107,6
2. Femmine disoccupate											
St. A	100	106,2	111,1	113,3	113,3	112,5	111,1	113,2	113,3	113,2	97,0
St. B	99,2	115,5	112,3	122,5	112,2	114,8	122,4		113,4	112,1	109,5
3. Popolazione occupata											
St. A	100	110,1	109,5	111,1	104,5	110,5	109,7	111,3	110,9	111,0	60,6
St. B	101,7	112,1	113,5	65,4	108,8	112,6	68,4		110,5	113,4	69,1
4. Femmine occupate											
St. A	100	113,3	109,5	111,1	109,1	110,7	109,7	111,4	112,5	111,1	99,0
St. B	102,4	119,0	114,5	108,6	113,1	115,0	110,2		114,5	114,4	110,5
5. Popolazione in condizione profess.											
St. A	100	119,5	110,6	112,3	107,1	112,3	110,8	112,8	111,6	112,3	56,1
St. B	101,8	121,5	115,1	65,9	111,2	117,5	67,5		112,2	115,0	64,3
6. Femmine in cond. profes.											
St. A	100	120,6	110,8	112,6	111,6	112,4	110,9	113,0	113,6	112,6	95,4
St. B	101,0	123,6	114,2	112,0	113,5	118,5	113,0		114,8	114,2	107,8
7. Residenti con scuola obbligo											
St. A	100	109,9	112,3	115,0	108,4	114,1	112,4	115,3	113,0	114,9	50,0
St. B	101,1	107,7	116,8	62,0	111,7	118,3	64,9		113,4	116,7	57,4
8. Femmine con scuola obbligo											
St. A	100	105,9	111,5	113,7	109,8	113,4	111,6	114,2	111,8	113,7	54,9
St. B	101,3	105,1	115,7	71,5	112,5	117,0	74,1		113,7	115,7	62,7
9. Popolazione con laurea											
St. A	100	129,1	117,7	123,8	123,9	119,5	117,3	122,4	120,8	123,7	85,8
St. B	96,9	124,3	117,6	124,2	117,8	140,2	124,3		120,2	117,4	104,4
10. Femmine laureate											
St. A	100	128,8	116,8	122,1	121,9	118,4	116,4	121,1	120,0	122,1	88,0
St. B	96,4	123,6	115,9	124,4	116,0	137,2	124,6		118,3	115,7	105,8

(a), (b), (c) e (d) vedi nota a pag. 131.

Il valore assoluto della varianza delle stime ottenute con la tecnica 1 è:

	M+F	F
	TOSCANA	
Popolazione disoccupata	1.001.062	406.357
Popolazione occupata	39.659.166	19.468.310
Popol. in cond. professionali	40.421.623	20.402.029
Residenti con scuola obbligo	14.073.912	3.633.315
Residenti con laurea	2.280.053	358.285
	PIEMONTE	
Popolazione disoccupata	1.027.082	354.690
Popolazione occupata	37.605.351	17.588.796
Popol. in cond. professionali	37.264.566	18.031.944
Residenti con scuola obbligo	28.066.795	8.361.082
Residenti con laurea	3.251.766	439.338
	CALABRIA	
Popolazione disoccupata	38.149.301,2	17.908.652,2
Popolazione occupata	41.860.024,5	21.284.080,3
Popol. in cond. professionali	53.741.579,9	38.876.872,1
Residenti con scuola obbligo	16.101.654,3	3.729.501,2
Residenti con laurea	2.163.698,1	492.522,5

Tavola 3 - Efficienza delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, per tecnica di stratificazione dei comuni, tecnica di campionamento dei comuni, stimatore adottato, variabile di riferimento e regione *(I: stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro; II: stratificazione per dimensione demografica dei comuni).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1		2		4		7		10	
	I	II	I	II	I	II	I	II	I	II
TOSCANA										
1. Popolazione disoccupata	100	97,8	131,2	134,8	146,7	146,9	126,2	137,6	146,3	145,3
2. Femmine disoccupate	100	99,4	130,7	134,7	143,3	144,0	125,7	137,8	143,0	143,1
3. Popolazione occupata	100	94,0	109,1	163,3	133,4	125,7	122,0	119,5	130,0	121,9
4. Femmine occupate	100	96,5	113,3	156,9	129,1	125,4	120,1	119,8	125,1	121,0
5. Popolazione in condizione profess.	100	94,4	112,4	164,9	132,3	126,4	121,9	120,1	128,7	122,4
6. Femmine in cond. profes.	100	97,0	116,5	157,0	129,3	126,7	120,4	120,9	125,1	122,0
7. Residenti con scuola obbligo	100	95,4	135,9	124,7	134,4	132,3	120,2	124,9	131,8	131,4
8. Femmine con scuola obbligo	100	97,5	143,3	124,5	130,8	129,9	119,5	124,2	129,4	129,4
9. Popolazione con laurea	100	107,9	150,8	192,0	132,5	150,9	116,5	139,7	131,8	150,4
10. Femmine laureate	100	111,3	156,9	204,2	133,2	156,0	116,8	144,0	131,8	155,9
PIEMONTE										
1. Popolazione disoccupata	100	162,3	119,5	204,4	116,1	198,9	107,1	195,0	109,5	198,9
2. Femmine disoccupate	100	167,9	117,7	198,5	115,0	205,9	106,9	201,2	109,3	206,0
3. Popolazione occupata	100	110,5	118,6	135,9	111,7	129,2	107,1	127,4	104,8	129,3
4. Femmine occupate	100	106,1	113,7	113,7	112,1	125,5	107,1	123,5	106,2	125,3
5. Popolazione in condizione profess.	100	114,4	123,0	139,2	111,6	133,3	107,1	131,7	102,7	133,5
6. Femmine in cond. profes.	100	108,6	119,1	115,8	112,3	128,0	107,1	126,3	104,9	127,9
7. Residenti con scuola obbligo	100	195,4	142,7	248,0	115,3	233,5	107,4	228,6	92,9	233,3
8. Femmine con scuola obbligo	100	190,9	137,1	247,1	115,6	225,5	107,4	222,0	96,6	225,2
9. Popolazione con laurea	100	71,5	146,5	96,1	133,5	92,7	109,2	89,6	126,8	92,6
10. Femmine laureate	100	92,1	154,4	120,5	130,4	118,5	108,3	114,2	122,9	118,4
CALABRIA										
1. Popolazione disoccupata	100	75,8	114,0	95,9	126,1	100,6	117,6	96,0	127,4	100,6
2. Femmine disoccupate	100	69,3	132,1	85,6	123,4	90,4	117,4	87,1	124,7	90,4
3. Popolazione occupata	100	107,0	108,4	120,6	125,0	133,5	115,6	130,8	119,9	133,5
4. Femmine occupate	100	97,5	119,2	110,9	129,4	121,8	115,5	119,5	127,1	121,8
5. Popolazione in condizione profess.	100	78,2	120,4	113,6	125,5	100,1	115,6	98,2	121,1	100,1
6. Femmine in cond. profes.	100	70,3	123,8	95,4	130,7	90,6	116,0	88,5	129,1	90,6
7. Residenti con scuola obbligo	100	103,3	131,4	166,8	134,9	139,6	115,7	133,0	134,3	139,5
8. Femmine con scuola obbligo	100	103,6	142,1	151,4	133,0	136,5	115,9	131,4	133,2	136,4
9. Popolazione con laurea	100	125,6	222,2	229,5	138,9	203,0	115,0	177,1	137,1	203,3
10. Femmine laureate	100	120,5	199,1	199,5	135,1	188,3	114,8	167,1	133,4	188,6

Il valore assoluto della varianza delle stime ottenute con la tecnica che è riportata in appresso (tra parentesi l'esponente della potenza con base 10 per cui moltiplicare il numero che lo precede).

	M+F	F
TOSCANA		
Popolazione disoccupata	0,6527942(6)	0,2693424(6)
Popolazione occupata	0,2486211(8)	0,1252921(8)
Popol. in cond. professionali	0,2544988(8)	0,1319680(8)
Residenti con scuola obbligo	0,8949971(7)	0,2362852(7)
Residenti con laurea	0,1640401(7)	0,2659029(6)
PIEMONTE		
Popolazione disoccupata	0,1111274(7)	0,3970308(6)
Popolazione occupata	0,2771183(8)	0,1244703(8)
Popol. in cond. professionali	0,2841978(8)	0,1305153(8)
Residenti con scuola obbligo	0,3656997(8)	0,1063946(8)
Residenti con laurea	0,1550083(7)	0,2697384(6)
CALABRIA		
Popolazione disoccupata	0,1929077(8)	0,8274346(7)
Popolazione occupata	0,2984854(8)	0,1383826(8)
Popol. in cond. professionali	0,2802818(8)	0,1822843(8)
Residenti con scuola obbligo	0,1109369(8)	0,2574729(7)
Residenti con laurea	0,1811037(7)	0,3956779(6)

Tavola 4 - Stabilità delle stime regionali della varianza riportate nella Tab. 1, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8	9	10	11
					TOSCANA					
1. Popolazione disoccupata										
St. A	100	183,9	158,3	145,2	266,5	112,4	240,1	255,8	76,6	3,5
St. B	219,9	157,9	207,0	198,3	170,6	212,6	208,9	203,6	186,3	4,7
2. Femmine disoccupate										
St. A	100	166,7	174,2	157,5	278,8	112,1	264,3	326,0	100,3	7,7
St. B	215,1	129,2	198,9	206,0	151,3	210,9	206,7	195,6	186,0	10,7
3. Popolazione occupata										
St. A	100	88,5	123,7	119,2	67,0	130,1	228,0	79,8	25,1	0,0
St. B	217,6	163,1	213,9	19,7	52,8	241,4	55,1	194,8	207,6	0,0
4. Femmine occupate										
St. A	100	77,4	122,1	118,5	236,5	138,7	241,7	163,3	40,6	0,1
St. B	210,6	152,0	204,7	142,0	193,0	239,4	171,8	200,0	198,3	0,2
5. Popolazione in condizione profess.										
St. A	100	95,5	122,8	118,6	73,6	128,7	248,1	81,9	26,4	0,0
St. B	231,0	183,0	226,1	21,0	56,5	250,4	59,2	205,3	218,7	0,0
6. Femmine in cond. profes.										
St. A	100	81,0	121,9	118,4	249,3	135,4	252,6	166,0	40,8	0,1
St. B	215,5	165,4	209,5	143,7	194,1	242,3	175,5	204,5	202,4	0,2
7. Residenti con scuola obbligo										
St. A	100	123,7	129,8	127,8	193,2	100,3	198,9	56,5	30,3	0,0
St. B	189,1	143,0	186,4	51,8	104,1	161,1	109,0	189,2	177,0	0,0
8. Femmine con scuola obbligo										
St. A	100	138,6	125,6	123,5	229,7	98,9	208,5	84,8	49,7	0,0
St. B	186,0	186,3	182,5	84,9	139,8	175,7	140,6	184,8	161,6	0,0
9. Popolazione con laurea										
St. A	100	105,7	138,3	138,4	370,0	118,7	335,3	332,7	90,0	8,1
St. B	147,6	77,0	138,5	142,0	147,4	138,2	142,3	140,9	136,7	10,4
10. Femmine laureate										
St. A	100	102,8	138,3	138,4	341,9	114,8	306,3	300,4	86,9	6,7
St. B	146,6	77,1	138,0	142,3	147,8	132,7	142,7	140,6	136,5	8,7

Tavola 4 segue - Stabilità delle stime regionali della varianza riportate nella Tab. 1, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8	9	10	11
PIEMONTE										
1. Popolazione disoccupata										
St. A	100	123,8	114,0	111,8	239,4	104,0	204,6	178,1	(a)	0,2
St. B	168,2	144,0	168,1	198,8	173,3	168,7	184,9	171,1	120,8	0,3
2. Femmine disoccupate										
St. A	100	116,4	107,7	105,0	323,0	104,5	283,1	253,7	128,6	0,5
St. B	201,5	198,2	203,2	230,2	194,9	198,3	215,8	200,7	120,5	0,5
3. Popolazione occupata										
St. A	100	79,6	103,4	99,0	38,7	106,2	266,5	46,4	(a)	0,0
St. B	272,2	378,4	274,2	19,2	19,3	275,8	158,4	262,9	105,2	0,0
4. Femmine occupate										
St. A	100	85,4	103,9	100,7	158,5	105,9	243,3	108,2	298,2	0,0
St. B	247,9	438,7	248,8	113,6	117,0	244,3	237,2	244,7	122,1	0,0
5. Popolazione in condizione profess.										
St. A	100	76,5	104,3	100,4	41,9	106,2	263,8	40,8	118,4	0,0
St. B	272,8	359,9	274,3	198,2	19,0	271,5	176,9	270,6	107,8	0,0
6. Femmine in cond. profes.										
St. A	100	80,6	104,8	101,8	165,0	105,8	231,9	95,1	69,3	0,0
St. B	249,0	426,9	249,5	112,3	113,7	243,4	244,9	248,9	122,3	0,0
7. Residenti con scuola obbligo										
St. A	100	80,0	110,9	110,8	365,9	104,6	213,1	79,8	(a)	0,0
St. B	268,4	374,1	269,3	212,6	215,4	265,2	346,1	299,5	160,5	0,0
8. Femmine con scuola obbligo										
St. A	100	79,2	109,6	109,1	353,8	104,4	212,4	94,5	823,3	0,0
St. B	251,6	444,0	252,5	232,3	231,5	243,2	306,4	270,0	162,9	0,0
9. Popolazione con laurea										
St. A	100	127,1	123,8	124,0	380,1	101,0	267,0	208,1	53,2	0,2
St. B	146,9	124,6	144,4	198,6	162,8	145,4	176,7	159,5	122,8	0,2
10. Femmine laureate										
St. A	100	146,1	120,5	121,0	338,8	99,7	249,8	200,6	88,2	0,2
St. B	146,0	126,7	142,6	207,2	161,3	143,9	176,8	159,0	117,9	0,2

(a) valore negativo

Tavola 4 segue - **Stabilità delle stime regionali della varianza riportate nella Tab. 1, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione** *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

Regione/ Variabile/ Stimatore	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8	9	10	11
CALABRIA										
1. Popolazione disoccupata										
St. A	100	104,3	125,3	116,0	266,7	139,6	298,1	346,3	55,1	10,1
St. B	246,4	235,0	239,9	236,0	194,1	212,9	238,7	230,3	224,0	13,0
2. Femmine disoccupate										
St. A	100	159,9	120,8	112,7	324,7	128,8	344,3	384,3	54,1	20,9
St. B	232,4	321,9	226,8	229,3	190,1	214,5	226,5	217,2	199,0	28,3
3. Popolazione occupata										
St. A	100	98,8	113,9	107,2	130,2	114,2	233,8	124,6	17,7	0,1
St. B	203,8	233,7	202,6	95,0	125,3	192,8	167,7	193,3	186,9	0,0
4. Femmine occupate										
St. A	100	123,9	127,3	122,6	157,8	121,6	192,7	202,1	21,4	0,6
St. B	169,2	189,2	167,7	141,9	145,0	162,4	168,2	166,1	159,5	0,8
5. Popolazione in condizione profess.										
St. A	100	87,5	112,9	106,8	115,1	139,4	320,2	159,0	25,5	0,0
St. B	207,7	293,0	208,7	59,1	97,2	218,7	122,4	193,4	185,0	0,0
6. Femmine in cond. profes.										
St. A	100	118,5	128,9	124,3	142,8	134,1	193,0	202,8	34,7	0,5
St. B	168,4	182,9	166,6	117,0	136,6	167,4	155,0	163,8	162,6	0,6
7. Residenti con scuola obbligo										
St. A	100	135,7	125,7	122,3	118,7	103,5	197,5	102,7	17,1	0,0
St. B	164,2	191,4	165,1	77,4	100,4	159,0	137,1	162,3	163,6	0,0
8. Femmine con scuola obbligo										
St. A	100	116,0	121,2	117,1	122,4	116,4	200,9	129,4	16,6	2,5
St. B	175,2	179,3	175,9	92,3	110,0	171,0	149,7	168,5	168,1	3,1
9. Popolazione con laurea										
St. A	100	163,0	156,5	155,6	342,1	107,3	258,7	225,3	34,6	1,1
St. B	145,3	82,1	140,6	157,4	152,5	136,9	156,4	147,0	143,1	1,4
10. Femmine laureate										
St. A	100	146,5	145,0	141,1	300,5	116,8	231,0	205,3	26,6	1,4
St. B	156,4	102,6	153,2	170,2	161,3	148,6	166,6	156,4	154,3	1,8

Il valore assoluto alla varianza delle stime ottenute con la tecnica 1 è riportata in appresso (tra parentesi l'esponente della potenza con base 10 per cui moltiplicare il numero che lo precede).

	M+F	F
	TOSCANA	
Popolazione disoccupata	0,1511991(12)	0,3745678(11)
Popolazione occupata	0,2000257(15)	0,6388485(14)
Popol. in cond. professionali	0,2357924(15)	0,7285059(14)
Residenti con scuola obbligo	0,2350179(14)	0,1900667(14)
Residenti con laurea	0,5899454(13)	0,1161917(12)
	PIEMONTE	
Popolazione disoccupata	0,3952647(12)	0,6818398(11)
Popolazione occupata	0,2117851(15)	0,4266239(14)
Popol. in cond. professionali	0,2256063(15)	0,4490158(14)
Residenti con scuola obbligo	0,4523257(15)	0,3743652(14)
Residenti con laurea	0,9020920(12)	0,2549639(11)
	CALABRIA	
Popolazione disoccupata	0,1979937(15)	0,5171189(14)
Popolazione occupata	0,2294126(15)	0,4238722(14)
Popol. in cond. professionali	0,2830108(15)	0,7010173(14)
Residenti con scuola obbligo	0,2831378(14)	0,1490342(13)
Residenti con laurea	0,2052547(13)	0,8557484(11)

Tavola 5 - Stabilità delle stime regionali della varianza riportate nella Tab. 2, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8	9	10	11
					TOSCANA					
1. Popolazione disoccupata										
St. A	100	180,9	160,6	154,6	276,9	158,9	246,9	304,4	115,8	19,3
St. B	228,9	244,2	210,8	209,8	180,6	205,0	212,2	208,6	207,6	25,7
2. Femmine disoccupate										
St. A	100	183,1	164,6	153,0	310,9	205,6	282,6	381,4	124,1	38,8
St. B	254,4	300,1	232,0	241,7	184,0	261,6	240,0	229,4	230,1	51,1
3. Popolazione occupata										
St. A	100	93,0	126,7	124,6	148,6	118,7	219,2	133,9	65,0	0,1
St. B	202,0	138,0	192,6	17,0	116,9	180,0	24,9	181,4	187,5	0,1
4. Femmine occupate										
St. A	100	96,5	126,5	125,2	240,0	126,7	216,5	188,1	77,4	0,8
St. B	182,7	132,2	172,7	117,0	180,9	174,1	131,0	169,6	167,9	0,9
5. Popolazione in condizione profess.										
St. A	100	95,9	125,8	123,8	158,3	112,5	233,3	137,1	67,4	0,1
St. B	202,4	145,2	193,2	169,1	118,8	178,7	24,7	182,0	188,1	0,1
6. Femmine in cond. profes.										
St. A	100	97,9	125,5	124,2	255,6	124,6	229,5	196,3	60,4	0,8
St. B	186,1	141,0	176,2	116,9	183,1	175,4	131,5	172,7	171,0	0,9
7. Residenti con scuola obbligo										
St. A	100	125,6	123,7	122,7	281,8	121,0	286,6	183,0	280,4	0,1
St. B	178,9	184,9	150,4	48,6	140,8	181,4	62,2	165,7	165,6	0,2
8. Femmine con scuola obbligo										
St. A	100	134,2	125,4	124,4	249,0	118,1	246,3	185,3	158,3	0,3
St. B	172,5	191,9	163,3	70,2	142,1	164,5	87,0	160,2	161,7	0,4
9. Popolazione con laurea										
St. A	100	209,9	149,6	150,6	306,5	158,2	264,9	272,0	155,2	33,3
St. B	164,4	188,1	151,1	147,3	169,2	159,5	148,4	152,4	151,1	51,2
10. Femmine laureate										
St. A	100	187,1	148,1	149,1	314,3	153,4	269,0	274,1	192,1	30,9
St. B	165,6	188,4	152,8	148,4	171,7	158,5	149,4	153,7	151,8	46,4

Tavola 5 segue - **Stabilità delle stime regionali della varianza riportate nella Tab. 2, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione** *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8	9	10	11
PIEMONTE										
1. Popolazione disoccupata										
St. A	100	106,1	114,4	113,9	290,8	104,6	279,1	295,6	(a)	55,8
St. B	156,3	156,0	151,7	150,0	148,7	145,4	150,2	151,3	144,3	64,9
2. Femmine disoccupate										
St. A	100	136,6	110,6	110,2	413,0	100,0	399,9	419,4	59,6	39,1
St. B	145,9	182,2	142,4	141,9	140,4	134,8	142,1	142,2	134,8	42,8
3. Popolazione occupata										
St. A	100	109,8	112,0	111,8	156,4	92,9	231,7	208,2	68,8	1,0
St. B	145,6	145,3	141,1	6,9	81,7	139,2	9,1	132,0	131,4	1,1
4. Femmine occupate										
St. A	100	99,2	115,6	115,4	213,9	105,3	214,9	209,4	78,4	6,3
St. B	133,5	116,0	129,0	54,3	123,8	125,4	62,5	128,3	127,0	7,6
5. Popolazione in condizione profess.										
St. A	100	104,7	111,6	111,3	155,7	103,0	241,6	214,4	123,7	0,9
St. B	148,7	142,0	144,2	6,1	79,6	141,5	8,1	133,6	132,6	1,1
6. Femmine in cond. profes.										
St. A	100	97,8	115,2	114,9	214,0	104,5	216,2	210,2	102,5	5,7
St. B	133,4	114,7	128,9	49,4	123,2	124,8	57,5	128,1	126,8	6,9
7. Residenti con scuola obbligo										
St. A	100	115,0	113,2	113,1	214,2	100,8	243,7	242,8	95,1	6,9
St. B	151,5	150,0	146,3	25,1	122,3	141,3	31,3	143,1	132,4	8,5
8. Femmine con scuola obbligo										
St. A	100	119,7	116,5	116,4	211,6	97,5	215,7	214,1	(a)	9,8
St. B	141,4	143,9	135,4	38,3	125,7	128,3	45,3	134,0	131,9	12,4
9. Popolazione con laurea										
St. A	100	157,7	145,1	145,6	507,8	78,9	460,2	493,2	144,5	69,8
St. B	157,8	156,1	142,8	141,2	147,9	124,3	141,6	143,3	142,7	110,2
10. Femmine laureate										
St. A	100	148,2	142,6	143,0	488,0	79,3	443,5	474,0	135,8	69,5

(a) Valore negativo.

Tavola 5 segue - Stabilità delle stime regionali della varianza riportate nella Tab. 2, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

Regione/ Variabile/ Stimatore	TECNICA DI CAMPIONAMENTO									
	1	2	3	4	6	7	8	9	10	11
					CALABRIA					
1. Popolazione disoccupata										
St. A	100	128,4	133,7	133,5	269,7	119,7	250,8	271,0	133,1	85,0
St. B	147,2	129,4	136,9	137,0	135,7	142,2	136,9	136,8	137,0	123,8
2. Femmine disoccupate										
St. A	100	108,8	131,4	131,3	345,2	106,0	320,4	343,9	131,1	80,1
St. B	142,0	109,9	133,3	133,0	134,0	130,0	133,0	133,4	133,3	113,4
3. Popolazione occupata										
St. A	100	123,2	123,6	123,1	232,0	106,0	239,2	253,0	120,7	31,4
St. B	140,2	131,2	132,3	74,0	122,9	134,2	81,9	131,7	132,0	43,9
4. Femmine occupate										
St. A	100	89,7	125,5	125,1	287,7	108,4	279,1	297,3	124,4	95,4
St. B	141,3	93,5	132,5	131,8	127,5	135,7	132,1	132,0	132,5	127,3
5. Popolazione in condizione profess.										
St. A	100	106,4	130,8	129,7	199,4	120,5	198,2	212,6	128,1	25,6
St. B	148,6	112,3	139,2	72,6	127,0	140,8	78,9	138,8	139,0	40,0
6. Femmine in cond. profes.										
St. A	100	91,5	134,7	133,8	192,9	115,1	180,9	196,4	133,2	74,2
St. B	150,8	94,7	140,9	139,5	137,7	139,8	140,6	140,6	140,9	110,5
7. Residenti con scuola obbligo										
St. A	100	115,7	136,7	135,8	185,7	94,7	186,1	198,8	135,6	22,9
St. B	151,7	125,3	142,6	64,3	129,0	127,1	71,2	142,1	142,5	38,1
8. Femmine con scuola obbligo										
St. A	100	120,8	134,3	133,6	185,5	100,5	181,4	184,1	132,8	29,2
St. B	150,5	131,5	141,4	87,0	131,0	131,0	93,4	141,3	141,3	44,5
9. Popolazione con laurea										
St. A	100	103,6	172,3	174,8	312,4	104,7	293,1	174,4	174,4	49,4
St. B	178,5	101,8	166,4	161,9	180,2	152,2	163,2	167,4	166,4	92,1
10. Femmine laureate										
St. A	100	93,8	157,7	159,1	335,0	106,1	273,5	320,8	158,7	51,3
St. B	164,7	91,0	152,9	149,7	164,4	142,3	150,7	154,1	152,8	85,9

Il valore assoluto della varianza delle stime ottenute con la tecnica 1 è riportata (tra parentesi l'esponente della potenza con base 10 per cui moltiplicare il numero che lo precede).

	M+F	F
TOSCANA		
Popolazione disoccupata	0,1337599(12)	0,3764034(11)
Popolazione occupata	0,1933458(15)	0,5696841(14)
Popol. in cond. professionali	0,2125721(15)	0,6339299(14)
Residenti con scuola obbligo	0,3167913(14)	0,2239262(13)
Residenti con laurea	0,3257142(13)	0,6520721(11)
PIEMONTE		
Popolazione disoccupata	0,9678457(11)	0,1699300(11)
Popolazione occupata	0,9643216(14)	0,1940288(14)
Popol. in cond. professionali	0,9720893(14)	0,2004949(14)
Residenti con scuola obbligo	0,5883290(14)	0,4527652(13)
Residenti con laurea	0,6746699(13)	0,7526626(11)
CALABRIA		
Popolazione disoccupata	0,2212510(15)	0,7208372(14)
Popolazione occupata	0,1780238(15)	0,5789307/14)
Popol. in cond. professionali	0,2574981(15)	0,1252880(15)
Residenti con scuola obbligo	0,2401075(14)	0,1291610(13)
Residenti con laurea	0,6541418(12)	0,4170920(11)

Tavola 6 - Stabilità delle stime regionali della varianza riportate nella Tab. 1, per tecnica di campionamento, stimatore adottato, variabile di riferimento e regione *(St. A: stimatore specifico; St. B: stimatore basato sul quoziente).*

REGIONE/ VARIABILE/ STIMATORE	TECNICA DI CAMPIONAMENTO							
	1		2		7		10	
	I	II	I	II	I	II	I	II
				TOSCANA				
1. Popolazione disoccupata	100	111,7	214,4	276,3	153,3	222,3	455,0	397,2
2. Femmine disoccupate	100	97,4	255,7	279,8	172,6	235,6	(a)	374,0
3. Popolazione occupata	100	104,6	123,8	103,2	137,2	145,1	175,9	111,7
4. Femmine occupate	100	115,6	151,1	117,5	144,7	153,4	115,8	136,3
5. Popolazione cond. prof.	100	112,7	125,8	115,9	151,3	318,3	188,7	226,8
6. Femmine in cond. prof.	100	117,9	148,1	121,9	148,3	311,1	115,5	250,2
7. Residenti scuola obbligo	100	67,3	161,3	113,4	110,6	131,6	235,1	95,1
8. Femmine con scuola obbligo	100	81,0	129,0	141,7	118,2	128,0	411,7	116,7
9. Popolazione con laurea	100	189,7	184,1	283,4	207,3	339,2	166,7	432,6
10. Femmine con laurea	100	182,8	180,5	265,1	185,8	333,0	166,3	408,7
				PIEMONTE				
1. Popolazione disoccupata	100	376,1	122,0	343,3	103,2	664,2	(a)	449,9
2. Femmine disoccupate	100	378,1	154,4	352,5	155,5	995,6	(a)	438,9
3. Popolazione occupata	100	228,8	94,8	262,9	146,2	304,3	0,6	(a)
4. Femmine occupate	100	228,9	116,6	304,2	131,2	282,7	0,8	872,8
5. Popolazione cond. prof.	100	238,4	98,1	272,2	144,5	333,9	0,7	(a)
6. Femmine in cond. prof.	100	229,9	114,1	302,8	123,5	285,6	0,7	603,8
7. Residenti con scuola obbligo	100	743,9	105,8	823,5	111,1	1079,1	0,5	985,2
8. Femmine con scuola obbligo	100	826,8	108,5	745,1	110,7	1031,7	1,2	1152,0
9. Popolazione con laurea	100	12,2	100,4	21,8	142,2	45,2	(b)	24,9
10. Femmine con laurea	100	30,5	93,3	30,5	130,0	107,0	0,9	59,9
				CALABRIA				
1. Popolazione disoccupata	100	93,8	166,0	155,6	187,4	157,5	170,4	164,6
2. Femmine disoccupate	100	73,2	133,5	114,7	220,9	164,6	(a)	122,1
3. Popolazione occupata	100	128,9	152,9	159,0	133,0	193,0	(a)	192,7
4. Femmine occupate	100	66,3	120,3	79,6	102,0	120,9	(a)	102,1
5. Popolazione cond. prof.	100	124,1	141,6	122,9	195,2	152,8	(a)	212,0
6. Femmine in cond. prof.	100	57,3	102,7	90,6	103,4	63,4	(a)	106,5
7. Residenti scuola obbligo	100	120,3	100,5	160,1	103,4	137,9	(a)	227,9
8. Femmine con scuola obbligo	100	119,8	109,9	164,4	107,1	130,5	(a)	222,2
9. Popolazione con laurea	100	315,5	128,5	871,4	144,0	599,1	(a)	898,9
10. Femmine con laurea	100	196,0	108,9	515,4	125,7	411,3	171,7	462,2

(a) Non si è riusciti a trovare uno stimatore specifico pratico dello stimatore basato sulla selezione senza reinserimento di Brewer (tecnica n. 4). - (b) Valore negativo.

Il valore assoluto della varianza delle stime ottenute con la tecnica 1 è riportato in appresso (tra parentesi l'esponente della potenza con base 10 per cui moltiplicare il numero che lo precede).

	M+F	F
TOSCANA		
Popolazione disoccupata	0,3594326(11)	0,9123443(10)
Popolazione occupata	0,4660248(14)	0,1510027(14)
Popol. in cond. professionali	0,1060015(15)	0,3218949(14)
Residenti con scuola obbligo	0,5240074(13)	0,4305578(12)
Residenti con laurea	0,1486763(13)	0,2879849(11)
PIEMONTE		
Popolazione disoccupata	0,8879832(11)	0,1668321(11)
Popolazione occupata	0,5132778(14)	0,1012003(14)
Popol. in cond. professionali	0,5453379(14)	0,1051933(14)
Residenti con scuola obbligo	0,1032777(15)	0,8536921(13)
Residenti con laurea	0,2173648(12)	0,6048502(10)
CALABRIA		
Popolazione disoccupata	0,4941929(14)	0,1322750(14)
Popolazione occupata	0,5394692(14)	0,0358196(13)
Popol. in cond. professionali	0,7121384(14)	0,1549842(14)
Residenti con scuola obbligo	0,6308923(13)	0,3345537(12)
Residenti con laurea	0,4904453(12)	0,1992519(11)

Tavola 7 - Caratteristiche delle stime regionali ottenute sommando stime sub-provinciali basate sulla selezione di due comuni per strato, per tecnica di campionamento e di stratificazione dei comuni, variabili di riferimento e regione (NC: non calcolato)

STRATIFICAZIONE TECNICA CAMPIONAMENTO	Numero di campioni di numerosità due (1)	Percentuale di campioni in cui compaiono probabilità negative (2)	Percentuale di strati in cui la tecnica non è correttamente applicabile (3)	Tempo-macchina per la esecuzione del programma (4)	Note (5)
			TOSCANA		
I. Stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro					
Tecnica 1	(b) 1.807	—		0	
2	252	—		0	
3	1.555	—		(d)	
4	3.110	—		(d)	π_{ij} è sempre inferiore a $\pi_i \pi_j$
5	1.555	1,4	31,8	(c)	
6	3.110	—	4,5	1	In uno strato l'algoritmo di calcolo delle probabilità di inclusione non converge. La somma delle probabilità di selezione è generalmente maggiore o uguale a 1 in ogni strato. π_{ij} è sempre inferiore a $\pi_i \pi_j$.
7	837	—		0	
8	3.110	—		(d)	La somma degli scarti quadratici (dal valore 1) delle probabilità di selezione è 2.032. π_{ij} è sempre inferiore a $\pi_i \pi_j$.
9	3.110	—			La somma degli scarti quadratici (dal valore 1) delle probabilità di inclusione è 2.996.
10	1.555	4,1	45,5	(d)	Lo spazio di memoria richiesto è un multiplo del numero di possibili campioni.
11	1.555			(d)	
II. Stratificazione per dimensione demografica					
Tecnica 1	(b) 2.216	—		1	
2	252	—		0	
3	1.961	—		(d)	
4	3.928	—		(d)	π_{ij} è sempre inferiore a $\pi_i \pi_j$.
5	1.964	—	22,7	(c)	π_{ij} è sempre inferiore a $\pi_i \pi_j$.
6	3.928	—		0	
7	1.043	—		0	
8	3.928	—		(d)	La somma degli scarti quadratici (dal valore 1) delle probabilità di selezione è 0,402. π_{ij} è sempre inferiore a $\pi_i \pi_j$.
9	3.928	—		(d)	La somma degli scarti quadratici (dal valore 1) delle probabilità di selezione è 0,541.
10	1.964	3,2	27,3	(d)	(v. I stratificazione)
11	1.964	—		(d)	

(a) Pur essendo calcolato in centesimi di secondo, il tempo viene approssimato al secondo nella stampa dell'output. Il tempo riportato nella colonna (4) è comprensivo anche dei tempi di gestione del programma da parte del sistema operativo. Il tempo «virtuale», ossia necessario all'elaborazione del programma, è spesso troncato perché non raggiunge il secondo. - (b) Il numero di campioni distinti è N (N — 1)/2, quello di campioni ripetuti è N. Gli uni e gli altri sono considerati nelle elaborazioni riportate nella presente tabella e altrove. - (c) Il tempo impiegato dalla tecnica n. 5 è ragionevolmente simile a quello della tecnica 7. - (d) Si può prendere come riferimento il tempo impiegato dalla tecnica n. 1.

Tavola 7 segue - **Caratteristiche delle stime regionali ottenute sommando stime subprovinciali basate sulla selezione di due comuni per strato, per tecnica di campionamento e di stratificazione dei comuni, variabili di riferimento e regione (NC: non calcolato)**

STRATIFICAZIONE TECNICA CAMPIONAMENTO	Numero di campioni di numerosità due (1)	Percentuale di campioni in cui compaiono probabilità negative (2)	Percentuale di strati in cui la tecnica non è correttamente applicabile (3)	Tempo-macchina per la esecuzione del programma (4)	Note (5)
			PIEMONTE		
I. Stratificazione derivata da quella in uso per l'indagine sulle forze di lavoro (solo 46 strati: 2 strati anomali sono ignorati)					
Tecnica 1	(b) 17.792	—		1	
2	1.163	—		0	
3	16.609	—		(d)	
4	33.218	—		(d)	π_{ij} è sempre inferiore a $\pi_i \pi_j$ nei 46 strati costruiti
5	16.609	0,6	15,2	(c)	
6	33.218	—	2,2	2	In uno strato l'algoritmo di stima delle probabilità non converge. La somma delle probabilità negli strati è generalmente superiore a 1. π_{ij} è sempre inferiore a $\pi_i \pi_j$ nei 46 strati.
7	8.589	—		1	
8	33.218	—		(d)	La somma degli scarti quadratici (dal valore 1) delle probabilità di selezione è 6.636. π_{ij} è sempre inferiore a $\pi_i \pi_j$ nei 46 strati costruiti.
9	33.218	—			La somma degli scarti quadratici (dal valore 1) delle probabilità di inclusione è 8.940.
10	16.609	10,1	76,1	(d)	Lo spazio di memoria richiesto è un multiplo del numero di possibili campioni.
11	16.609			(d)	
II. Stratificazione per dimensione demografica (48 strati)					
Tecnica 1	(d) 30.150	—		1	
2	1.179	—		1	
3	28.971	—		(d)	
4	57.942	—		(d)	π_{ij} è sempre inferiore a $\pi_i \pi_j$
5	28.971	—	14,7	(c)	π_{ij} è sempre inferiore a $\pi_j \pi_j$.
6	57.942	—		3	
7	14.775	—		2	
8	57.942	—		(d)	La somma degli scarti quadratici (dal valore 1) delle probabilità di selezione è 0,260. π_{ij} è sempre inferiore a $\pi_i \pi_j$.
9	57.942	—		(d)	La somma degli scarti quadratici (dal valore 1) delle probabilità di selezione è 0,453.
10	28.971	0,9	8,3	(d)	(v. I stratificazione)
11	289.714	—		(d)	

(a) Pur essendo calcolato in centesimi di secondo, il tempo viene approssimato al secondo nella stampa dell'output. Il tempo riportato nella colonna (4) è comprensivo anche dei tempi di gestione del programma da parte del sistema operativo. Il tempo «virtuale», ossia necessario all'elaborazione del programma, è spesso troncato perché non raggiunge il secondo. - (b) Il numero di campioni distinti è N (N — 1)/2, quello di campioni ripetuti è N. Gli uni e gli altri sono considerati nelle elaborazioni riportate nella presente tabella e altrove. - (c) Il tempo impiegato dalla tecnica n. 5 è ragionevolmente simile a quello della tecnica 7. - (d) Si può prendere come riferimento il tempo impiegato dalla tecnica n. 1.

NOTE

(*) L'Autore desidera ringraziare i membri della «Commissione per la progettazione e l'applicazione dei campioni» istituita presso l'Istat per i preziosi suggerimenti ricevuti sia nella fase di impostazione delle elaborazioni, sia in quella di analisi dei risultati descritti nella presente memoria, nonché la prof.a Daniela Cocchi per le considerazioni conseguenti ad una prima versione dello scritto. Un riconoscimento particolare va, inoltre, ai dott. M. Gaggiotti e A. Zuchegna dell'Istat per aver svolto con pazienza e competenza le elaborazioni informatiche.

Di ogni giudizio e degli eventuali errori è responsabile unicamente l'Autore.

(1) A quello proposto da Horvitz e Thompson (1952) è preferito, come è prassi comune, lo stimatore della varianza di Sen-Yates-Grundy perché questo, pur non godendo delle stesse proprietà teoriche, è decisamente superiore in pratica, generando pochi valori di varianza negativi (Rao e Singh, 1973).

(2) Prima di rendere operative le indicazioni che seguono questi discorsi è conveniente estendere la sperimentazione all'intera Italia, per variabili diverse (non solo variabili di livello, ma anche medie, rapporti e altre statistiche non lineari) e con altre alternative di stratificazione.

(3) Lo stimatore basato sul quoziente è stato applicato solo per $n = 2$.

(4) Rao e Bayless (1969), esaminando campioni di 2 unità tratte da 7 popolazioni artificiali e da 20 insiemi naturali di unità, trovano che lo stimatore di Raj (1956) e quello proposto da Murthy (1957) sono praticamente equivalenti quanto a stabilità. Tuttavia, non sempre lo stimatore più efficiente (quello di Murthy) si mostra anche il più stabile. Bayless e Rao (1970), per $n = 3$, trovano che lo stimatore di Murthy è appena più stabile di quello di Raj.

RIFERIMENTI BIBLIOGRAFICI

BAYLESS D.L. e RAO J.N.K. (1970) An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4), *Journal of the American Statistical Association*, 65: 1645-1667.

BEALE E.M.L. (1962) Some use of computers in operational research, *Industrielle organisation*, 31, 27-28.

BIGGERI L., CHIANDOTTO B. e GHILARDI G. (1977) *Materiale di discussione dei primi risultati di una stratificazione dei comuni della Toscana*, Commissione per gli studi statistici ed econometrici interessanti la programmazione economica, Documento n. 58, Istat, Roma.

BREWER K.R.W. (1963) A model of systematic sampling with unequal probabilities, *Australian Journal of Statistics*, 5: 5-13.

BREWER K.R.W. (1975) A simple procedure for sampling pswor, *Australian Journal of Statistics*, 17, 166-172.

BREWER K.R.W. (1979) A class of robust sampling designs for large-scale surveys, *Journal of the American Statistical Association*, 74: 911-915.

BREWER K.R.W. e HANIF M. (1969) Sampling without replacement with probability of inclusion proportional to size. I: Methods using Horvitz-Thompson estimator (Manoscritto non pubblicato).

BREWER K.R.W. e HANIF M. (1983) *Sampling with Unequal Probabilities*, Springer-Verlag, New York.

CARROLL J.L. e HARTLEY H.O. (1964) The symmetric method of unequal probability sampling without replacement. Abstract in *Biometrics*, 20: 908-909.

COCHRAN W.G. (1953; 1963; 1977) *Sampling Techniques*, Wiley, New York.

DALENIUS T. (1957) *Sampling in Sweden, Contributions to the Theory of Sample Survey Practice*, Almqvist & Wiksell, Stockholm.

DALENIUS T. e GURNEY M. (1951) The problem of optimum stratification. II, *Skandinavisk Aktuarietidskrift*, 34: 133-148.

DURBIN J. (1967) Design of multistage surveys for the estimation of sampling errors, *Applied Statistics*, 16: 152-164.

FABBRIS L. (1981) Alcune proposte sul tema del sovracampionamento su base regionale e provinciale dell'indagine sulle forze di lavoro, *Economia & Lavoro*, anno XV, n. 3: 19-34.

FABBRIS L. (1983) Problemi statistici per il sovracampionamento su base regionale del campione nazionale delle forze di lavoro. In: Trivellato U. e Zuliani A. (a cura di) *Informazione statistica su scuola e mercato del lavoro e sulle politiche per l'occupazione giovanile*, Min. P.I., Istituto della Enciclopedia Italiana, Roma: 235-257.

FABBRIS L. (1985) Problemi inerenti alla selezione di due comuni per strato nelle indagini sulla popolazione in Italia (dattiloscritto), Istat, Roma.

FELLEGI I.P. (1963) Sampling with varying probabilities without replacement: rotating and non-rotating samples, *Journal of the American Statistical Association*, 58: 183-201.

GODAMBE V.P. (1960) An admissible estimate for any sampling design, *Sankhyā (A)*, 22: 286-288.

GODAMBE V.P. e JOSHI V.M. (1965) Admissibility and Bayes estimation in sampling finite populations, I, II and III, *Annals of Mathematical Statistics*, 36: 1707-1742.

GOODMAN L. e KISH L. (1950) Controlled selection - a technique in probability sampling, *Journal of the American Statistical Association*, 45: 350-372.

GOSH D.N. (1983) Determining the sample size for multivariate estimates satisfying simultaneous confidence intervals, *Proceedings of the Section on Survey Research Methods, American Statistical Association*: 727.

HANIF M. e BREWER K.R.W. (1980) Sampling with unequal probabilities without replacement: a review, *International Statistical Review*, 48: 317-335.

HANSEN M.H. e HURWITZ W.N. (1943) On the theory of sampling from finite populations, *Annals of Mathematical Statistics*, 14: 333-362.

HANSEN M.H., HURWITZ W.N. e MADOW W.G. (1953) *Sample Survey Methods and Theory*, Wiley, New York.

HANURAV T.V. (1968) Hyper-admissibility and optimum estimators for sampling from finite population, *Annals of Mathematical Statistics*, 39: 621-642.

HARTLEY H.O. (1962) Sampling with unequal probabilities and without replacement, *Annals of Mathematical Statistics*, 33: 3650-374.

HARTLEY H.O. (1966) Systematic sampling with unequal probability and without replacement, *Journal of the American Statistical Association*, 61: 739-748.

HARTLEY H.O. e RAO J.K.W. (1962), Sampling with unequal probabilities and without replacement, *Annals of Mathematical Statistics*, 33: 350-374.

HARTLEY H.O., RAO J.N.K., e KIEFER G. (1969) Variance estimation with one unit per stratum, *Journal of the American Statistical Association*, 64: 841-851.

HARTLEY H.O. e ROSS A. (1954) Unbiased ratio estimators, *Nature*, 174: 270-271.

HERZEL A. (1984) Campionamento senza ripetizione con probabilità diverse: piani di campionamento con probabilità di inclusione di qualunque ordine preassegnato. In: *Atti della XXXII Riunione Scientifica, vol. I, Società Italiana di Statistica, Sorrento, 11-13 aprile 1984*: 211-225.

HESS I., SRIKANTAN K.S. (1970) Recommended variables for the multiple stratification of general hospitals, *Health Service Research*, 5: 12-24.

HORVITZ D.G. e THOMPSON D.J. (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47: 663-685.

ISTAT (1958, 1969, 1978) *Rilevazioni campionarie delle forze di lavoro*, Metodi e Norme, Serie A, nn. 3, 10, 15.

KISH L. (1965) *Survey Sampling*, Wiley, New York.

KISH L. (1976) Optima and proxima in linear sample design, *Journal of the Royal Statistical Society (A)*, 139: 80-95.

MADOW W.G. (1949) On the theory of systematic sampling, II, *Annals of Mathematical Statistics*, 20: 333-354.

MAHALANOBIS P.C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute (e discussione), *Journal of the Royal Statistical Society*, 109: 325-370.

MURTHY M.N. (1957) Ordered and unordered estimators in sampling without replacement, *Sankhyā*, 18: 379-390.

National Commission on Employment and Unemployment Statistics (1979) *Counting the Labor Force*, Washington, DC.

PATHAK P.K. (1966) An estimator in PPS sampling for multiple characteristics, *Sankhyā (A)*, 28: 35-40.

PATHAK P.K. (1967a) Asymptotic efficiency of the symmetrized Des Raj strategy, II, *Sankhyā (A)*, 29: 299-304.

PATHAK P.K. (1967b) Asymptotic efficiency of Des Raj's strategy, I, *Sankhyā, (A)* 29: 283-298.

QUENOUILLE M.H. (1956) Notes on bias in estimation, *Biometrika*, 43: 353-360.

RAJ D. (1956) Some estimators in sampling with varying probabilities without replacement, *Journal of the American Statistical Association*, 51: 269-284.

RAJ D. (1958) On the accuracy of some sampling techniques, *Journal of the American Statistical Association*, 53: 98-101.

RAO J.N.K. (1963) On two systems of unequal probability sampling without replacement, *Annals Inst. Statist. Association*, 3: 173-180.

RAO J.N.K. (1966) On the relative efficiency of some estimators in PPS sampling for multiple characteristics, *Sankhyā (A)*, 28: 61-70.

RAO J.N.K. e BAYLESS D.L. (1969) An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum, *Journal of the American Statistical Association*, 64: 540-549.

RAO J.N.K., HARTLEY H.O. e COCHRAN W.G. (1962) On a simple procedure of unequal probability sampling without replacement, *Journal of the Royal Statistical Society (B)*, 24: 482-491.

RAO J.N.K. e SINGH M.P. (1973) On the choice of estimators in survey sampling, *Australian Journal of Statistics*, 15: 95-104.

RAO P.S.R.S. e RAO J.N.K. (1971) Small sample results for ratio estimators, *Biometrika*, 58: 625-630.

RAO T.J. (1967) On the choice of a strategy for the ratio metod of estimation, *Journal of the Royal Statistical Society (B)*, 29: 392-397.

ROY J. e CHAKRAVARTI I.M. (1960) Estimating the mean of a finite population, *Annals of Mathematical Statistics*, 31: 392-398.

ROYALL R.M. e CUMBERLAND W.G. (1981) An empirical study of the ratio estimator and estimators of its variance, *Journal of the American Statistical Association*, 76: 66-78.

RUSSO A. e FALORSI P. (1985) Rilevazioni campionarie delle forze di lavoro. Metodologia del campionamento, calcolo e presentazione errori campionari. Quaderni di discussione, Istat, Roma.

SAMPFORD M.R. (1967) On sampling without replacement with unequal probabilities of selection, *Biometrika*, 54: 499-513.

SEN A.R. (1953) On the estimate of the variance on sampling with varying probabilities, *Journal of the Indian Society of Agricultural Statistics*, 5: 119-127.

SHAPIRO G.M. e OLSEN C.L. (1979) Should one or two PSU's per stratum be selected, *Proceedings of the Section on Survey Research Methods*, American Statistical Association: 314-318.

SUNTER A.B. (1977) Response burden, sample rotation, and classification renewal in economic surveys, *International Statistical Review*, 45: 209-222.

SUNTER A.B. (1986) Solution to the problem of unequal probability sampling without replacement, *International Statistical Review*, 54: 33-50.

TADROS W.H., MOORE T.F. e CHAKRABARTY R.P. (1982) Determining the optimal numbers of Primary Sampling Units to be selected for the Health Interview Survey, *Proceedings of the Section on Survey Research Methods*, American Statistical Association: 217-222.

TIN M. (1965) Comparison of some ratio estimators, *Journal of the American Statistical Association*, 60: 294-307.

YATES F. e GRUNDY P.M. (1953) Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society (B)*, 15: 253-261.

ZANI S. e SICURI S. (1977a) Primi risultati di una stratificazione dei comuni della Emilia-Romagna sulla base di indicatori socio-economici, Commissione per gli studi statistici ed econometrici interessanti la programmazione economica, Documento n. 56, Istat, Roma.

ZANI S. e SICURI S. (1977b) Stratificazione dei comuni dell'Emilia-Romagna con l'impiego del metodo gerarchico e del metodo non gerarchico, Commissione per gli studi statistici ed econometrici interessanti la programmazione economica, Documento n. 59, Istat, Roma.

ZANNELLA F. (1982), Piano di rilevazione campionaria ed errori di campionamento. In: De Sandre P. (a cura di) *Indagine sulla fecondità in Italia. 1979. Rapporto Generale, Vol. I: Metodologia e Analisi*, Tecnoprint, Bologna: 49-71.

ARCHITETTURA DELLA PROCEDURA GENERALIZZATA PER LA STRATIFICAZIONE E LA SELEZIONE DEI COMUNI NELLE INDAGINI CAMPIONARIE SULLA POPOLAZIONE

di *L. Fabbris e F. Zannella*

1. INTRODUZIONE

Nella presente nota viene riportato uno schema di riferimento per mettere a punto una procedura informatica per la stratificazione e la selezione di un campione di comuni italiani.

A regime, la procedura dovrebbe consentire all'utilizzatore di estrarre, in tempi brevi e con semplici istruzioni, un campione di comuni da un archivio magnetico seguendo un disegno campionario a due stadi con stratificazione delle unità di primo stadio e loro selezione con probabilità variabile. La procedura, che si denomina «generalizzata» per le flessibilità introdotte, consente anche il calcolo delle varianze campionarie delle stime di frequenze relative (o assolute) di alcune caratteristiche della popolazione rilevate al censimento e assimilate a variabili oggetto d'indagine.

Le informazioni che rendono attivo il programma sono costituite da:

a) un file di dati comunali, contenente per ciascun comune i valori delle variabili desunte dal movimento anagrafico e dal censimento della popolazione del 1981, aggiornati alla data più recente di disponibilità dei dati del movimento anagrafico;

b) un file contenente per ciascun dominio territoriale la numerosità del campione, i costi unitari previsti in ciascuno stadio di campionamento e il numero minimo e massimo d'interviste d'assegnare a ciascun rilevatore;

c) la data di riferimento dell'indagine espressa in giorno, mese ed anno;

d) alcuni parametri standard, scelti dagli autori della presente nota, che consentono il funzionamento del programma anche in assenza di scelte metodologiche da parte dell'utente.

L'output è dato da 4 files per i quali è prevista la stampa su carta e/o la registrazione su disco, in modo da consentirne ulteriori elaborazioni:

1. il file dei comuni universo, nel quale per ogni comune sono riportati la denominazione, i codici identificativi (provincia e comune all'interno della provincia), i codici del dominio territoriale e dello strato di appartenenza, la popolazione e il numero di famiglie residenti alla data di riferimento dell'indagine, i valori delle variabili scelte come obiettivo ai fini della programmazione del campione;

2. il file degli strati, contenente per ciascuno strato i codici del dominio territoriale e dello strato, la popolazione residente (totale, media, minima e massima), il numero di famiglie (totale, medio, minimo e massimo) e alcune statistiche descrittive (media, minimo, massimo, varianza, coefficiente di variazione) delle variabili obiettivo;

3. il file dei comuni campione in cui, per ogni comune selezionato, sono riportate tutte le informazioni contenute nel file dei comuni universo e il numero di unità di secondo stadio che devono essere campionate;

4. il file dei domini territoriali, contenente per ciascun dominio il codice identificativo, il numero dei comuni universo, la popolazione e il numero di famiglie universo, il numero degli strati e dei comuni campione distinti in autorappresentativi e non, la numerosità del campione in secondo stadio.

Va precisato che le dimensioni del campione al primo e al secondo stadio devono essere imputati e non sono un risultato delle elaborazioni previste dalla procedura. Tuttavia, dopo aver esaminato i valori delle varianze campionarie calcolate per ognuna delle variabili obiettivo, si possono apportare aggiustamenti alle numerosità e ripetere la procedura immettendo le nuove dimensioni del campione.

Dal punto di vista informatico, la procedura dovrebbe essere costituita da un insieme di programmi gestiti da comandi EXEC e pannelli DMS per la scelta delle opzioni previste.

Nel grafico 1 è illustrato lo schema generale della procedura, mentre nel paragrafo 2 sono riportati i pannelli con le opzioni previste nelle varie fasi.

Nel paragrafo 3 sono descritte schematicamente e commentate le tecniche di selezione dei comuni scelte per essere inserite nella procedura.

2. DESCRIZIONE DEI PANNELLI PER LA SCELTA DELLE OPZIONI

Pannello PA1: imputazione della data di riferimento dell'indagine

giorno
mese
anno

N.B. invio = richiesta completata
PF3 = richiesta annullata

Pannello PA2: scelta dei domini territoriali

codice del dominio territoriale
nome del dominio territoriale

ITALIA NORD-OCCIDENTALE

Piemonte	Valle d'Aosta	Lombardia
Liguria		

ITALIA NORD-ORIENTALE

Trentino-A.A.	Veneto	Friuli-V.G.
Emilia-Romagna		

ITALIA CENTRALE

Toscana	Umbria	Marche
Lazio		

ITALIA MERIDIONALE

Abruzzi	Molise	Campania
Puglia	Basilicata	Calabria

ITALIA INSULARE

Sicilia	Sardegna

N.B. 1 = intera ripartizione o intera regione
2 = scelta delle regioni entro le ripartizioni o delle province entro le regioni
invio = richiesta completata
PF3 = richiesta annullata

Pannello PA3: compilazione delle schede parametro

Compaiono tante schermate quanti sono i domini territoriali richiesti, ciascuna contenente il codice e il nome del dominio e le variabili per le quali devono essere forniti i valori:

codice del dominio territoriale
nome del dominio territoriale
numero di comuni campione
numero di comuni da estrarre per strato
tasso di campionamento finale
numero minimo d'interviste per rilevatore
numero massimo d'interviste per rilevatore
costo previsto per unità di primo stadio
costo previsto per unità di secondo stadio
soglia di popolazione per le delimitazione
dei comuni autorappresentativi

N.B. il codice e il nome del dominio sono riportati da programma, le altre informazioni devono essere digitate dall'utilizzatore; la soglia di popolazione non va digitata quando per la determinazione dei comuni autorappresentativi si sceglie uno dei metodi automatici previsti in PA5.

invio = imputazione completata
PF3 = imputazione annullata

Pannello PA4: scelta delle variabili oggetto di studio

Superficie
Popolazione residente
Popolazione residente per sesso
Popolazione residente per stato civile
Popolazione residente per classi quinquennali di età
Popolazione residente per particolari classi di età
Popolazione residente da 6 anni in poi per titolo di studio
Popolazione residente che frequenta corsi regolari di studio o corsi di formazione professionale
Popolazione residente attiva per condizione
Popolazione residente attiva in condizione professionale per posizione nella professione
Popolazione residente attiva in condizione professionale per ramo di attività economica
Popolazione residente non attiva per classi di età e condizione
Famiglie residenti e componenti

Abitazioni occupate
Abitazioni occupate per titolo di godimento
Abitazioni occupate per servizio installato
Indicatori socio-demografici

N.B. 1 = tutte le modalità del carattere
2 = scelta di una o più modalità
invio = richiesta completata
PF3 = richiesta annullata

Pannello PA5: scelta del metodo per la determinazione della soglia dei comuni autorappresentativi

soglia variabile digitata in PA3
numero minimo d'interviste per rilevatore
metodo di Hidiroglou
soglia costante per tutti i domini

N.B. per scegliere uno dei primi tre metodi digitare 1 accanto al metodo scelto se si vuole scegliere il quarto metodo digitare accanto il valore della soglia

invio = richiesta completata
PF3 = richiesta annullata

Pannello PA6: scelta delle variabili di stratificazione

Superficie
Popolazione residente
Popolazione residente per sesso
Popolazione residente per stato civile
Popolazione residente per classi quinquennali di età
Popolazione residente per particolari classi di età
Popolazione residente da 6 anni in poi per titolo di studio
Popolazione residente che frequenta corsi regolari di studio o corsi di formazione professionale
Popolazione residente attiva per condizione
Popolazione residente attiva in condizione professionale per posizione nella professione
Popolazione residente attiva in condizione professionale per ramo di attività economica
Popolazione residente non attiva per classi di età e condizione

Famiglie residenti e componenti
Abitazioni occupate
Abitazioni occupate per titolo di godimento
Abitazioni occupate per servizio installato
Indicatori socio-demografici
Componenti principali

N.B. 1 = tutte le modalità del carattere
2 = scelta di una o più modalità
invio = richiesta completata
PF3 = richiesta annullata

Pannello PA7: scelta del metodo per la stratificazione dei comuni non autorappresentativi

A. Metodi per una sola variabile di stratificazione
uguale numero di comuni per strato
uguale ammontare della variabile di stratificazione
cumulata della radice quadrata della frequenza
cluster analysis
imputazione degli estremi delle classi

B. Metodi per due o più variabili di stratificazione
modalità prevalente
cluster analysis
imputazione degli estremi delle classi

N.B. va digitato 1 accanto al metodo scelto
invio = richiesta completata
PF3 = richiesta annullata

Pannello PA8: imputazione del giudizio sulla stratificazione ottenuta

Sei soddisfatto della stratificazione ottenuta?

Sì
No

N.B. va digitato 1 accanto alla risposta scelta
invio = richiesta completata
PF3 = richiesta annullata

Pannello PA9: scelta del metodo per la selezione dei comuni.

A. Metodo per n = 1

selezione con probabilità proporzionale all'ampiezza

B. Metodi per n = 2

tecnica sistematica di Madow-Hartley
tecnica di selezione senza reimmissione di Brewer
tecnica dei substrati casuali di Rao-Hartley-Cochran
tecnica di selezione di campioni interi di Herzel

C. Metodi per n > 2

tecnica sistematica di Madow-Hartley
tecnica di selezione senza reimmissione di Brewer
tecnica dei substrati casuali di Rao-Hartley-Cochran

N.B. va digitato 1 accanto al metodo scelto
invio = richiesta completata
PF3 = richiesta annullata

Pannello PA10: inputazione del giudizio sui risultati ottenuti

Sei soddisfatto dei risultati ottenuti?

Sì
No

N.B. va digitato 1 accanto alla risposta scelta
invio = richiesta completata
PF3 = richiesta annullata

3. LE TECNICHE DI SELEZIONE UTILIZZABILI

Qualora l'utente non esprima una scelta precisa tra le tecniche implementate, è imposta automaticamente la procedura di Madow-Hartley. Questa consiste nella selezione sistematica del campione con probabilità proporzionale alla dimensione delle unità. La selezione avviene dopo che le unità sono state accoppiate per formare

tanti possibili campioni quante sono le unità della popolazione secondo la procedura descritta in Hartley (1966):

Il vantaggio della tecnica sta nella semplicità del procedimento e nell'essere la selezione sistematica intuitiva per qualsiasi utente. È, inoltre, una procedura che dà stime efficienti, in modo particolare se a monte le unità sono state stratificate per dimensione, e con varianza di campionamento stabile, anche se si adottano stimatori asintotici.

Si precisa innanzi tutto che, quando l'utente specifica di voler selezionare più di un comune per strato, conviene che la numerosità campionaria sia 2, per i seguenti motivi:

(i) controllo profondo nella selezione delle unità campionarie

(ii) massimo guadagno potenziale in termini di efficienza delle stime grazie all'estensione della stratificazione

(iii) disponibilità di stimatori specifici per un gran numero di tecniche di selezione del campione

(iv) se si pensa di adottare la tecnica di compenetrazione delle assegnazioni degli intervistatori, due è il minimo numero di comuni per strato per stimare separatamente l'errore campionario e l'errore di rilevazione e che minimizza le spese di viaggio dei rilevatori.

Le tecniche selezionate per essere applicate nella procedura CAMPOP sono:

A) Per n = 2 una tra le tecniche seguenti: tecnica di Madow-Hartley (Grafico 2), tecnica di Rao-Hartley-Cochran (Grafico 4), selezione di campioni interi di Herzel (Grafico 5), in quanto tutte le tecniche:

(i) sono facilmente implementabili su calcolatore;

(ii) generano campioni di unità distinte;

(iii) generano campioni con probabilità positive (se si applica alla tecnica di Herzel (1984) la correzione dallo stesso proposta) e le probabilità di selezione sono facilmente calcolabili;

(iv) sono efficienti, le varianze sono stabili più del campionamento casuale semplice per quasi tutte le variabili e se co-

niugate con la stratificazione per dimensione, i risultati attesi sono considerevolmente efficienti.

B) Per n > 2, una tra le tecniche seguenti: tecnica di Madow-Hartley selezione senza reimmissione di Brewer (Grafico 3), tecnica di Rao-Hartley-Cochran in quanto:

(i) tutte le tecniche sono implementabili sul calcolatore senza eccessive difficoltà. La tecnica di Rao-Hartley-Cochran ammette varianti rispetto alla procedura indicata in Fabbris (1985);

(ii) tutte le tecniche generano campioni con unità distinte;

(iii) tutte le tecniche generano campioni con probabilità positive;

(iv) tutte le tecniche sono altamente efficienti e le varianze delle stime sono stabili più del campionamento con reimmissione. Per campioni di numerosità superiori a 3, non è data una forma esatta per il calcolo delle probabilità di selezione con la tecnica senza reimmissione proposta da Brewer. Se coniugate con la stratificazione per dimensioni dei comuni, i risultati attesi sono di notevole efficienza.

Grafico 1. Diagramma di flussi tra blocchi di operazioni per l'utilizzazione della procedura CAMPOP

Con il comando exec CAMPOP viene mandata in esecuzione le procedura generalizzata che prevede l'accesso al file aggiornato dei dati comunali e il richiamo di una serie di programmi (PR) e di pannelli (PA) per la scelta delle opzioni previste o per l'immissione delle informazioni richieste.

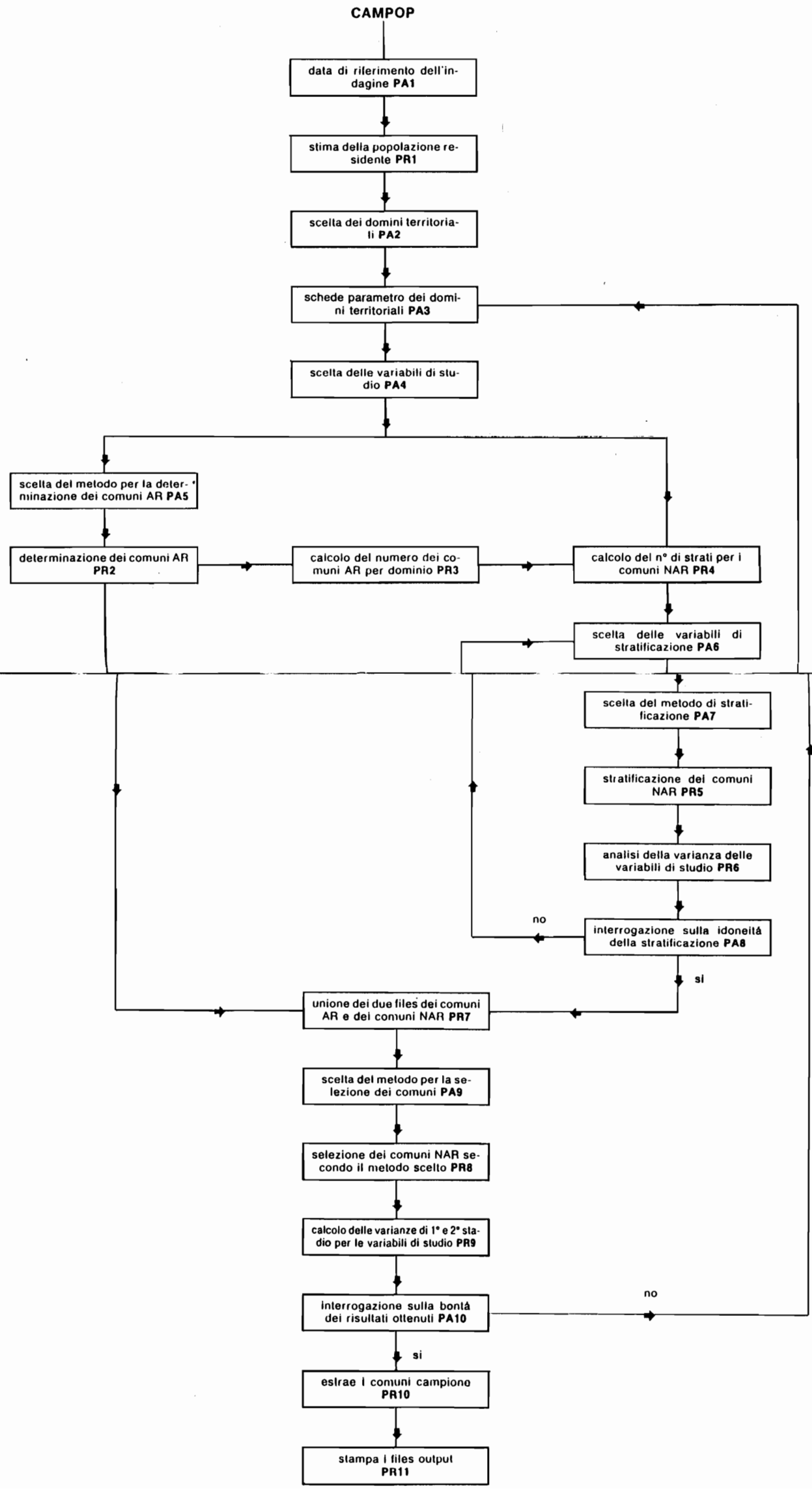


Grafico 2. Diagramma dei flussi tra blocchi di operazioni per l'applicazione della tecnica sistematica di Madow - Hartley

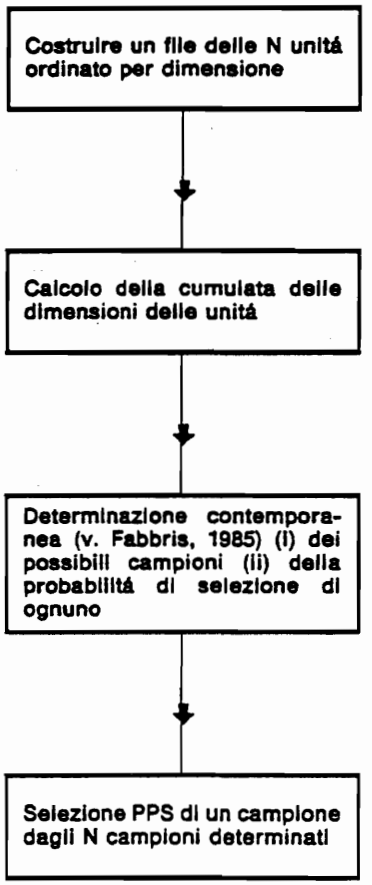


Grafico 3. Diagramma dei flussi tra blocchi di operazioni per l'applicazione della tecnica di selezione senza reimmissione di Brewer

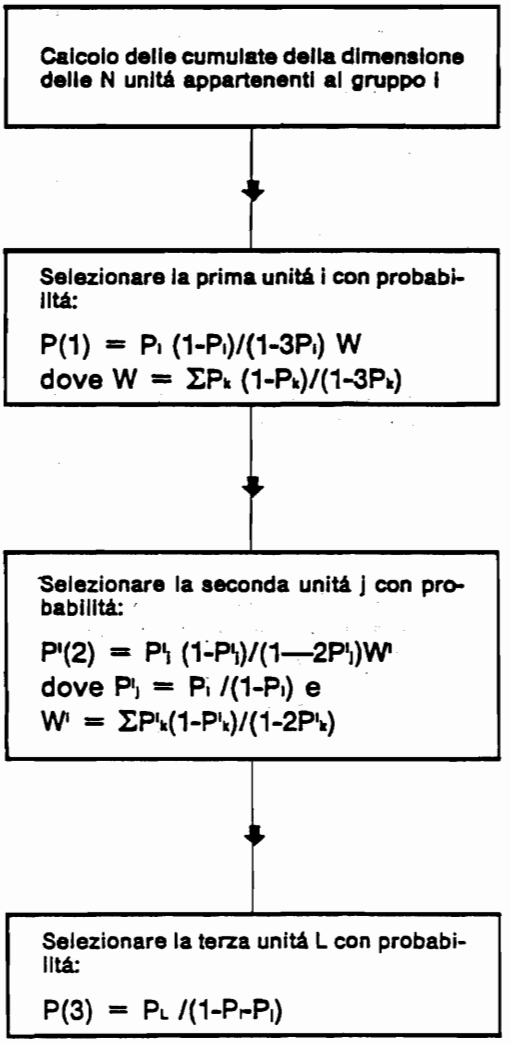


Grafico 4. Diagramma dei flussi tra blocchi di operazioni per l'applicazione della tecnica dei substrati casuali di Rao - Hartley - Cochran

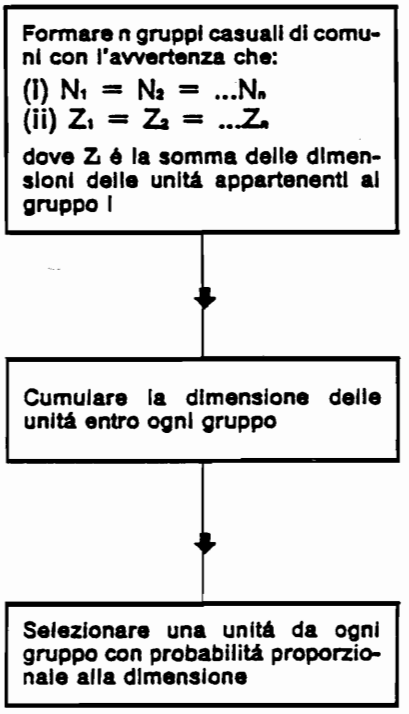
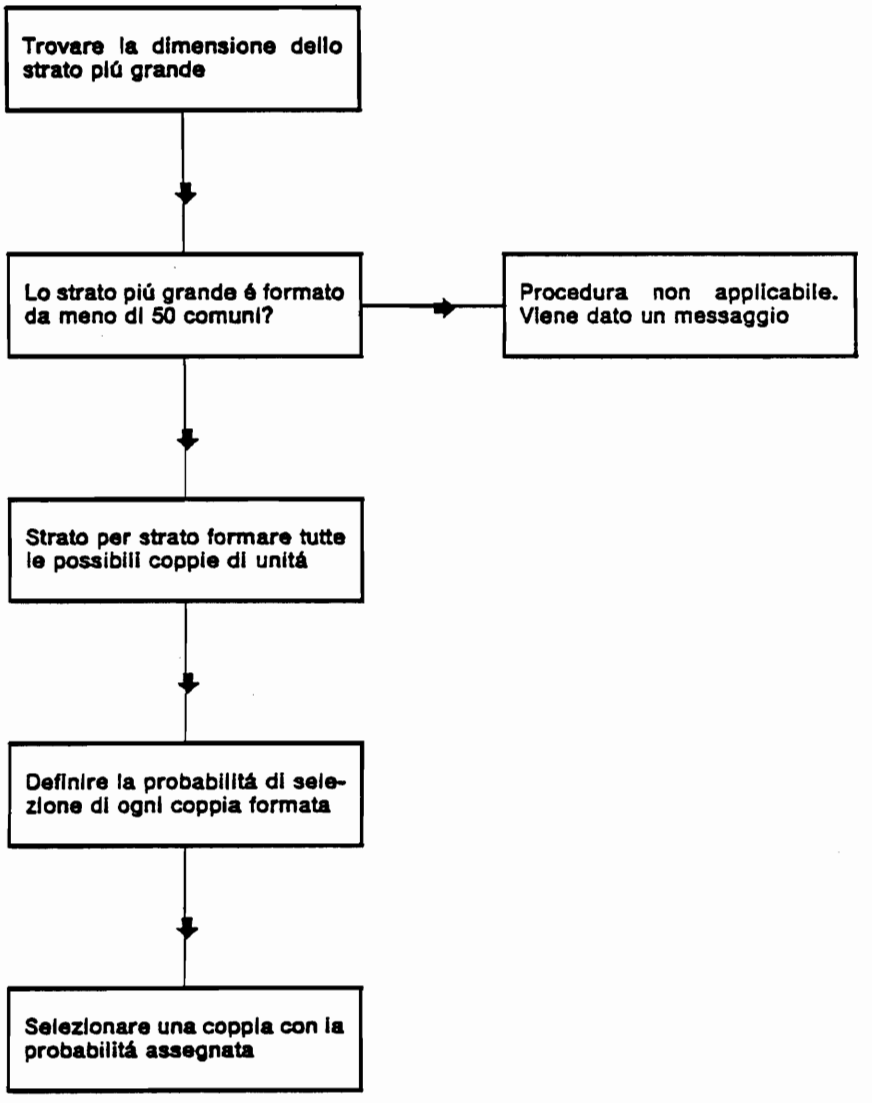


Grafico 5. Diagramma dei flussi tra blocchi di operazioni per l'applicazione della tecnica di selezione di campioni interi di Herzel.



STIMATORI UTILIZZATI NELLE INDAGINI ISTAT CONDOTTE SULLE FAMIGLIE: CONTRIBUTI METODOLOGICI E PRINCIPALI RISULTATI EMPIRICI

di *Stefano Falorsi*

1. INTRODUZIONE

Un problema che merita particolare attenzione nella predisposizione di una strategia campionaria è quello relativo alla scelta dello stimatore da adottare per l'ottenimento delle stime delle caratteristiche della popolazione oggetto di indagine.

Con riferimento alle indagini effettuate dall'Istat, la soluzione di tale problema presenta aspetti metodologici ed operativi di natura diversa intimamente legati al contesto campionario e alle informazioni ausiliarie disponibili.

A questo riguardo, come è noto, il complesso delle indagini Istat si può suddividere nei seguenti due gruppi fondamentali [1]:

— indagini sulle famiglie;
— indagini sulle aziende.

Per le indagini del primo gruppo si ricorre a disegni di campionamento a due stadi con stratificazione delle unità primarie; per le rimanenti indagini, nella maggior parte dei casi, vengono adottati disegni ad uno stadio stratificato.

In questa nota illustriamo gli aspetti teorici ed i risultati empirici più significativi delle ricerche condotte sulla tematica degli stimatori, nel corso di questi ultimi anni. Saranno altresì illustrate le implicazioni che da tali ricerche sono scaturite ai fini della scelta degli stimatori utilizzati, e verranno svolte alcune considerazioni conclusive sia sui problemi ancora aperti sia sulle direttrici future di ricerca.

L'illustrazione riguarderà soltanto le indagini sulle famiglie, in quanto rappresentano il terreno sul quale sono stati sviluppati in modo più intenso i contributi in tema di stimatori [2].

2. STIMATORI UTILIZZATI NEL PASSATO

Prima di illustrare le finalità e gli aspetti salienti degli studi più recenti, riteniamo utile passare brevemente in rassegna i metodi di stima utilizzati nel corso di questi ultimi anni.

All'inizio degli anni '80, per tutte le indagini condotte sulle famiglie veniva utilizzato il campione dell'indagine sulle Forze di Lavoro; in alcuni casi si ricorreva all'intero campione, in altri, invece, veniva adottato un sub-campione (3).

In particolare per l'ottenimento delle stime oggetto di indagine veniva usato uno stimatore del rapporto separato post-stratificato per sesso; riteniamo utile precisare che tale stimatore è tuttora utilizzato per l'indagine sulle Forze di Lavoro (4).

La struttura formale dello stimatore in oggetto è definita dall'espressione:

$$\hat{Y}_{ps} = \sum_{a=1}^{2} \sum_{h \in H} \frac{{}_a\hat{Y}_h}{{}_a\hat{P}_h} {}_aP_h \qquad (1)$$

in cui: ${}_a\hat{Y}_h$ = stima corretta del numero di persone, di sesso a, residenti nello strato h, che presentano la generica caratteristica y; ${}_a\hat{P}_h$ = stima corretta del numero di persone di sesso a, residenti nello strato h; ${}_aP_h$ = popolazione di sesso a residente nello strato h.

La (1) può riscriversi nella forma più semplice:

$$\hat{Y}_{ps} = \sum_{a=1}^{2} \sum_{h \in H} \frac{{}_ay_h}{{}_ap_h} {}_aP_h \qquad (2)$$

in cui ${}_ay_h$ = numero di persone intervistate, di sesso a, nello strato h, che presentano la caratteristica y; ${}_ap_h$ = numero di persone intervistate, di sesso a, nello strato h.

Lo stimatore (2), consente di elevare il livello di precisione rispetto a quello ottenibile mediante l'uso dello stimatore diretto (basato esclusivamente sull'utilizzazione delle probabilità di inclusione), gode anche della proprietà che la struttura della popolazione per sesso al livello di strato, stimata attraverso il campione, risulta uguale a quella della popolazione residente nel medesimo strato (5).

3. PRIMI CONTRIBUTI

Dai primi anni '80 si è iniziata a modificare la politica seguita nella predisposizione delle indagini campionarie Istat; in particola-

re ci si è resi autonomi dal piano di campionamento utilizzato per le Forze di Lavoro e si è adottata la logica di programmare strategie campionarie ad hoc, in funzione degli obiettivi conoscitivi di ciascuna indagine [6].

Una prima esperienza in tal senso è costituita dall'«Indagine sulle strutture e i comportamenti familiari» (Russo e Di Traglia, 1985).

Nell'indagine in oggetto è stato adottato uno stimatore del rapporto combinato espresso da:

$$\hat{Y}_c = \frac{\hat{Y}_d}{\hat{P}_d} P \qquad (3)$$

dove P indica la popolazione residente nel dominio di riferimento; $\hat{Y}_d$ e $\hat{P}_d$, rappresentano le stime dirette rispettivamente del totale della caratteristica di interesse e del totale della popolazione. Tali stime sono fornite dalle due seguenti relazioni:

$$\hat{Y}_d = \sum_{h \in H} \sum_{i \in n_h} \sum_{j \in m_{hi}} K_{hij} Y_{hij} \qquad (4)$$

$$\hat{P}_d = \sum_{h \in H} \sum_{i \in n_h} \sum_{j \in m_{hi}} K_{hij} P_{hij} \qquad (5)$$

in cui h rappresenta l'indice di strato; i l'indice di comune, j l'indice di famiglia; H insieme degli strati definiti nel dominio territoriale; n_h insieme dei comuni campione nello strato h; m_{hi} insieme delle famiglie campione nel comune (hi); K_{hij} peso diretto della famiglia (hij); Y_{hij} numero di componenti della famiglia (hij) che presentano la caratteristica di interesse; P_{hij} numero di componenti della famiglia (hij).

Nell'indagine in oggetto in ciascuna ripartizione geografica sono stati individuati due domini territoriali di riferimento: il primo costituito dai comuni con popolazione superiore a 100.000 abitanti; il secondo costituito dai rimanenti comuni.

Dalle relazioni (3), (4) e (5) segue immediatamente che il peso finale attribuito alla generica famiglia (hij) è dato da:

$$K_{hij} \frac{P}{\hat{P}_d} \qquad (6)$$

Lo stimatore in esame, che si basa sull'utilizzazione della popolazione come variabile ausiliaria, verifica la condizione che la stima della popolazione totale riferita al generico dominio risulta uguale all'ammontare della popolazione residente del dominio. Inoltre, si ritiene utile sottolineare che il peso espresso dalla (6), in considerazione del fatto che tutti i componenti della famiglia estratta vengono intervistati rappresenta anche il peso assegnato a ciascuno di essi. Tale peculiarità, ha costituito, in un certo senso, un elemento di cui si è tenuto conto nel momento della scelta dello stimatore da adottare per l'ottenimento delle stime dell'indagine, che aveva la finalità di fornire stime sia delle caratteristiche delle famiglie che degli individui.

In tali circostanze, il fatto di poter disporre di un unico sistema di pesi assicura, come illustreremo con maggiore dettaglio nel seguito, l'uguaglianza tra le stime desumibili delle tavole di pubblicazione relative agli individui con quelle ottenibili dalle tavole riportanti le caratteristiche delle famiglie.

La stessa finalità si poteva raggiungere utilizzando uno stimatore diretto; si è preferito tuttavia ricorrere allo stimatore del rapporto combinato in quanto esso assicura stime più precise ed allo stesso tempo, come abbiamo sopra sottolineato, garantisce il rispetto della condizione di uguaglianza tra l'ammontare noto della popolazione residente e quello ottenuto in base alla stima campionaria.

Per l'«Indagine sulla lettura e su altri aspetti del tempo libero, anno 1984», (Russo, Falorsi, Coccia e Giovani, 1986) è stato adottato uno stimatore del rapporto combinato e post-stratificato per sesso, espresso da:

$$\hat{Y}_{pc} = \sum_{a=1}^{2} \frac{{}_a\hat{Y}}{{}_a\hat{P}} \, {}_aP \qquad (7)$$

in cui a (= 1,2) indica il sesso; ${}_aP$, la popolazione residente di sesso a nel dominio di riferimento; ${}_a\hat{Y}$ e ${}_a\hat{P}$ costituiscono le stime rispettivamente del totale della caratteristica di interesse e del totale della popolazione, relativamente al sesso a nel dominio di riferimento; tali stime sono espresse da:

$${}_a\hat{Y} = \sum_{h \in H} \sum_{i \in n_h} \sum_{j \in m_{hi}} K_{hij} \, {}_aY_{hij} \qquad (8)$$

$$ {}_a\hat{P} = \sum_{h \in H} \sum_{i \in n_h} \sum_{j \in m_{hi}} K_{hij}\, {}_aP_{hij} \qquad (9) $$

dove ${}_aY_{hij}$ indica il numero di componenti, di sesso a della famiglia (hij), che presentano la caratteristica di interesse e ${}_aP_{hij}$ indica il numero di componenti di sesso a della famiglia (hij).

Nell'indagine in oggetto nell'ambito di ciascuna ripartizione geografica sono stati individuati due domini di riferimento: il primo costituito dai comuni con popolazione superiore a 100.000 abitanti, il secondo costituito dai rimanenti comuni.

Dalle relazioni precedenti si ricava che il peso finale attribuito a tutti i componenti di sesso a della famiglia (hij) è dato da:

$$ K_{hij} \frac{{}_aP}{{}_a\hat{P}} \qquad (10) $$

Lo stimatore espresso dalla (7) presenta la proprietà di garantire l'uguaglianza tra l'ammontare noto della popolazione residente per sesso e quello ottenuto in base alle stime campionarie. L'impiego di tale stimatore per l'indagine in questione, per la quale non si aveva la necessità di attribuire un unico sistema di pesi per famiglie ed individui, fu dettato dalla considerazione che si voleva da una parte aumentare l'efficienza delle stime suddividendo il campione in gruppi più omogenei secondo il sesso, e dall'altra si voleva evitare il rischio di avere distorsioni elevate conseguenti alle basse numerosità campionarie dei singoli strati elementari.

4. RECENTI SVILUPPI

4.1 Stimatore del rapporto combinato post-stratificato per sesso e classi di età

In occasione della «Indagine sugli sport e sulle vacanze, nel 1985», (Falorsi, Coccia e Russo 1988; Falorsi, Coccia, Russo e Botta, 1988), si è presa la decisione di usare uno stimatore del rapporto combinato post-stratificato per sesso e classi di età. La struttura formale di tale stimatore è espressa dalla seguente relazione:

$$\hat{Y}_{pc} = \sum_{a \in A} \frac{{}_a\hat{Y}}{{}_a\hat{P}} \, {}_aP \qquad (11)$$

dove a indica la classe demografica (le classi demografiche sono individuate dalle modalità congiunte dei caratteri sesso e classe di età [7]; A è l'insieme delle classi demografiche, ${}_aP$ indica la popolazione appartenente alla classe demografica a nel dominio di riferimento (regione geografica), risultante dalle statistiche demografiche; ${}_a\hat{P}$ e ${}_a\hat{Y}$ indicano rispettivamente, con riferimento alla classe demografica a e al generico dominio, la stima diretta del totale della popolazione e la stima diretta del totale della caratteristica di interesse. Queste ultime stime possono ottenersi mediante espressioni analoghe alle (8) e (9).

Lo stimatore appena illustrato consente di ottenere l'uguaglianza tra la distribuzione della popolazione regionale residente per sesso e classi di età e quella stimata attraverso il campione; inoltre conduce alla determinazione di stime più affidabili rispetto a quelle ottenibili attraverso l'uso di stimatori del tipo già descritto. La maggior efficienza di questo stimatore è da ascriversi al fatto che esso si basa sulla suddivisione del campione di unità finali in sub-popolazioni (definite combinando le modalità del sesso e dell'età) che risultano molto omogenee rispetto alle variabili oggetto di studio. Tale proprietà trova conferma anche sul piano empirico, così come hanno mostrato alcune ricerche (Russo, 1988; Russo e Botta, 1989) finalizzate allo studio comparativo dei principali stimatori adottati nelle indagini Istat.

Gli stimatori presi in considerazione sono:

— lo stimatore diretto, definito dalla (4);

— lo stimatore del rapporto combinato, espresso dalla relazione (3);

— lo stimatore del rapporto combinato post-stratificato per sesso e classi di età, fornito dalla (11).

Il confronto tra i diversi stimatori è stato effettuato mediante il calcolo dei seguenti indici di efficienza:

$$\hat{E}_{c,d} = \left(\frac{\hat{V}(\hat{Y}_c)}{\hat{V}(\hat{Y}_d)} \right)^{0,5} \qquad (12)$$

in cui $\hat{E}_{c,d}$ rappresenta l'efficienza dello stimatore combinato rispetto allo stimatore diretto, $\hat{V}(\hat{Y}_d)$ e $\hat{V}(\hat{Y}_c)$ indicano le stime delle varianze campionarie rispettivamente di $\hat{Y}_d$ e $\hat{Y}_c$;

$$\hat{E}_{pc,c} = \left(\frac{\hat{V}(\hat{Y}_{pc})}{\hat{V}(\hat{Y}_c)} \right)^{0,5} \tag{13}$$

in cui $\hat{E}_{pc,c}$ rappresenta l'efficienza dello stimatore del rapporto combinato post-stratificato rispetto allo stimatore del rapporto combinato in cui $\hat{V}(\hat{Y}_{pc})$ indica, la stima della varianza campionaria di $\hat{Y}_{pc}$;

$$\hat{E}_{pc,d} = \left(\frac{\hat{V}(\hat{Y}_{pc})}{\hat{V}(\hat{Y}_d)} \right)^{0,5} \tag{14}$$

in cui $\hat{E}_{pc,d}$ rappresenta l'efficienza dello stimatore del rapporto combinato post-stratificato rispetto allo stimatore diretto.

Nei lavori citati sono illustrate le espressioni utilizzate per il calcolo delle stime delle varianze in oggetto.

La sperimentazione è stata condotta con riferimento all'«Indagine sugli sport e sulle vacanze», anno 1985, limitatamente alle regioni Valle d'Aosta e Toscana.

Nella tabella seguente sono riassunti i risultati della ricerca.

Tabella 1 - Efficienza degli stimatori del rapporto combinato e del rapporto combinato post-stratificato rispetto allo stimatore diretto

Stime N.	$\hat{E}_{c,d}$	$\hat{E}_{pc,c}$	$\hat{E}_{pc,d}$
VALLE D'AOSTA			
Occupati	0,78	0,80	0,62
Persone andate al mare	0,81	0,93	0,75
Persone andate in albergo	0,99	0,97	0,96
Persone andate in vacanza	0,81	0,94	0,76
TOSCANA			
Occupati	0,59	0,82	0,47
Persone andate al mare	0,79	0,96	0,76
Persone andate in albergo	0,89	0,98	0,87
Persone andate in vacanza	0,76	0,98	0,75

Dalla Tab. 1 si evince, secondo le attese, che lo stimatore meno efficiente è quello diretto, peraltro da noi considerato solo come stimatore di riferimento. Lo stimatore che, in assoluto, consente di ottenere i guadagni più elevati rispetto a quello diretto è lo stimatore del rapporto combinato post-stratificato per sesso e classi di età: il campo dei valori di $\hat{E}_{pc,d}$ varia da un minimo di 0,47 a un massimo di 0,96. Guadagni di un certo rilievo, sempre con riferimento allo stimatore diretto, si ottengono con lo stimatore del rapporto combinato.

4.2 Procedure Raking

Un problema che si presenta nelle indagini Istat è quello di utilizzare un metodo di stima che:

i) consenta di rispettare la condizione di uguaglianza tra i totali noti della popolazione e le corrispondenti stime campionarie: più precisamente, tali totali nel caso delle Indagini Istat condotte sulle famiglie — sono costituiti dai valori delle sub-popolazioni definite dal concatenamento delle modalità del sesso e di quelle relative alle classi di età;

ii) assicuri una perfetta coerenza tra le stime relative alle famiglie e quelle relative agli individui.

Allo scopo di chiarire quest'ultimo punto consideriamo le seguenti tabelle:

Tabella 2 - Distribuzione delle famiglie secondo il numero di componenti occupati

1	...	*i*	...	s
$\hat{F}_1$	...	$\hat{F}_i$	...	$\hat{F}_s$

Tabella 3 - Distribuzione degli occupati secondo il titolo di studio

Elementare	Media Inferiore	Media Superiore	Laurea
$\hat{N}_1$	$\hat{N}_2$	$\hat{N}_3$	$\hat{N}_4$

Ciò premesso, nel caso in cui si attribuisce un primo sistema di pesi alle famiglie campione (per l'ottenimento delle stime $\hat{F}_i$, i = 1, .., s) e un secondo sistema agli individui (per l'ottenimento delle stime $\hat{N}_l$, l = 1, .., 4) si ha che la stima «numero complessivo degli occupati» desumibile dalla Tab. 2 è in generale diversa da quella desumibile dalla Tab. 3.

Per superare tale inconveniente, come abbiamo già sottolineato precedentemente, è necessario utilizzare un unico sistema di pesi, nel senso che a ciascuna famiglia campione e a ciascun componente della stessa deve essere assegnato lo stesso peso.

Una prima soluzione al problema in questione è stata data in connessione con l'«Indagine sulle condizioni di salute degli italiani; anno 1987».

La procedura di stima descritta nelle sue linee metodologiche essenziali in Falorsi (1988), si basa sulla risoluzione di un problema di minimo vincolato.

L'insieme dei vincoli è costituito da un sistema di A equazioni, in cui la generica equazione relativa alla generica classe demografica a [8], è espressa da:

$$\sum_{h\in H} \sum_{i\in n_h} \sum_{j\in m_{hi}} {}_aP_{hij}\, K_{hij}\, T_{hij} = {}_aP \qquad (a\in A) \qquad (15)$$

in cui T_{hij} (h = 1, ..., H; i = 1, ..., n_h; j = 1, ... m_{hi}) rappresentano le incognite; il sistema espresso nella (15) pertanto, coinvolge m incognite, dove $m = \sum_{h\in H} \sum_{i\in n_h} m_{hi}$ indica il numero di famiglie campione in un dato dominio di riferimento (ad esempio, regione geografica).

Prima di proseguire nell'illustrazione del metodo riteniamo opportuno ricordare che le quantità ${}_aP_{hij}\, K_{hij}$ e ${}_aP$ sono note e che le incognite T_{hij} in sostanza rappresentano dei coefficienti di correzione dei pesi diretti K_{hij}.

Al fine di descrivere in forma semplice e sintetica è conveniente introdurre la seguente simbologia:

$\underline{B} = \{{}_aP_{hij}\, K_{hij}\}$ è una matrice ad A righe ed m colonne;

$\underline{P} = \{{}_aP\}$ è il vettore colonna A-dimensionale dei termini noti;

$\underline{T} = \{T_{hij}\}$ è il vettore colonna m-dimensionale dei correttori da determinare.

Il sistema (15) espresso in termini matriciali è dato da:

$$\underline{B}\,\underline{T} = \underline{P} \tag{16}$$

Se indichiamo con r il rango di $\underline{B}$ il sistema ammette ∞^{m-r} soluzioni possibili; tra queste, la soluzione da noi adottata, è stata determinata imponendo l'ulteriore vincolo definito dalla relazione seguente:

$$\min \left[(\underline{T} - \underline{1})^t (\underline{T} - \underline{1}) \right] \tag{17}$$

in cui $\underline{1}$ rappresenta il vettore m-dimensionale costituito da tutti 1.

La soluzione del sistema di minimo vincolato, definito dalla (16) e dalla (17), ottenuta mediante il metodo dei moltiplicatori di Lagrange è data da:

$$\underline{T} = \underline{B}^t (\underline{B}\,\underline{B}^t)^{-1}\,\underline{P} - \underline{B}^t (\underline{B}\,\underline{B}^t)^{-1}\,\underline{B1} + \underline{1} \tag{18}$$

Il peso finale relativo a tutti i componenti della famiglia (hij) è pertanto dato da:

$$S_{hij} = K_{hij}\, T_{hij} \tag{19}$$

in conseguenza, la stima del totale della caratteristica di interesse nel dominio di riferimento è data da:

$$\hat{Y}_R = \sum_{h \in H} \sum_{i \in n_h} \sum_{j \in m_{hi}} S_{hij}\, Y_{hij} \tag{20}$$

Riteniamo opportuno sottolineare che nel sistema non è introdotto il vincolo $T_{hij} > 0$, per ogni $h \in H$, $i \in n_h$ e $j \in m_{hi}$; pertanto il metodo in questione non assicura che i pesi finali S_{hij} siano tutti positivi.

Le esperienze da noi condotte su questo terreno hanno mostrato che il numero di pesi negativi è tanto più elevato quanto più è elevata la discrepanza tra la struttura nota e quella ottenibile mediante l'utilizzazione dei pesi originari (diretti).

Tuttavia, il metodo in questione, adottato per l'indagine sulle condizioni di salute, ha portato a risultati molto soddisfacenti in quanto su circa 25.000 pesi soltanto tre sono risultati negativi.

Per superare tale inconveniente sono state studiate in profondità alcune procedure iterative che garantiscono l'ottenimento di soluzioni che conducono a pesi positivi.

Nel prosieguo illustreremo gli aspetti metodologici essenziali della procedura ritenuta più valida e che sarà utilizzata per la determinazione delle stime dell'indagine «Multiscopo».

Il sistema sul quale si basa tale procedura è costituito dall'insieme di equazioni definite dalle espressioni:

$$\sum_{h\epsilon H} \sum_{i\epsilon n_h} \sum_{j\epsilon m_{hi}} {}_aP_{hij}\, W_{hij} = {}_aP \qquad (a\epsilon A) \tag{21}$$

dove W_{hij} indica il peso finale relativo alla famiglia (hij). La soluzione del sistema (W_{hij}; $h\epsilon H$, $i\epsilon n_h$, $j\epsilon m_{hi}$) viene determinata sotto la condizione di minimizzazione della seguente funzione obiettivo:

$$\sum_{h\epsilon H} \sum_{i\epsilon n_h} \sum_{j\epsilon m_{hi}} P_{hij}\, W_{hij}\, \ln (W_{hij} / K_{hij}) \tag{22}$$

È chiaro che, essendo, la funzione obiettivo non lineare, per risolvere il sistema individuato dalle (21) e (22) è necessario ricorrere a un criterio iterativo.

Il criterio iterativo adottato che è descritto nel lavoro di Alexander e Roebuch (1986), parte dalla soluzione iniziale data da:

$$W_{hij}\,(O) = K_{hij}\,(P/\hat{P}_d) \tag{23}$$

la soluzione relativa alla t-esima iterazione è fornita dalla seguente espressione:

$$W_{hij}\,(t) = W'_{hij} \prod_{a\epsilon A} \left[{}_aP \,/ \sum_{h\epsilon H} \sum_{i\epsilon n_h} \sum_{j\epsilon m_{hi}} ({}_aP_{hij}\, W'_{hij}) \right]^{({}_aP_{hij} \,/\, P_{hij})} \tag{24}$$

in cui per ragioni di spazio si è posto $W'_{hij} = W_{hij}\,(t - 1)$.

Le sperimentazioni condotte utilizzando le informazioni dell'indagine Multiscopo hanno mostrato che la procedura in oggetto risulta caratterizzata da una velocità di convergenza che riteniamo accettabile.

Al fine di individuare l'iterazione in corrispondenza della quale arrestare il processo iterativo si è posta la condizione che lo scarto relativo tra il generico valore noto e il corrispondente ottenuto con

la procedura fosse inferiore all'1‰; tale risultato è stato conseguito con un numero di iterazioni, variabile da regione a regione, e in ogni caso inferiore a 50 iterazioni.

Allo scopo di eliminare le piccole discrepanze tra l'ammontare complessivo della popolazione residente e quello ottenuto mediante la procedura si è ritenuto opportuno introdurre un coefficiente di correzione determinato con la procedura (non iterativa) già descritta.

5. CONSIDERAZIONI CONCLUSIVE E FUTURI ITINERARI DI RICERCA

Nei paragrafi precedenti abbiamo svolto alcune considerazioni di natura concettuale e metodologica concernenti le tendenze più recenti in tema di stimatori utilizzati per l'effettuazione delle indagini Istat.

Lo stimatore del rapporto combinato post-stratificato e la procedura raking, anche se non rappresentano un traguardo di arrivo, costituiscono a nostro parere un ulteriore passo in avanti nell'ambito di un filone di ricerca il cui obiettivo è la determinazione di stime più affidabili, ovviamente a parità di condizioni.

Relativamente a tali procedimenti di stima sono stati già avviati alcuni studi che hanno la finalità di mettere a punto delle procedure informatiche generalizzate e flessibili in modo da poterle utilizzare in contesti campionari diversi.

Delle due procedure riteniamo di dover privilegiare quella raking in quanto la storia passata e le tendenze attuali fanno ritenere che le indagini da effettuare sulle famiglie sono orientate a produrre stime di caratteristiche relative sia agli individui che alle famiglie.

Un altro obiettivo di ricerca, che riteniamo di notevole importanza, è quello riguardante la definizione di uno stimatore composto, da adottare nelle indagini periodiche.

Su questo terreno sono stati già avviati studi, con riferimento all'indagine sulle forze di lavoro, sulle questioni concernenti il problema della rotazione intimamente connesso con la definizione della struttura formale dello stimatore composto.

NOTE

(1) È opportuno precisare che l'Istat effettua su particolari sub-popolazioni altre indagini nelle quali, però, non interviene la famiglia come unità di campionamento (ad esempio, l'«Indagine nella distribuzione per età della popolazione scolastica» e l'«Indagine sullo sbocco professionale dei laureati»).

(2) Nell'ambito delle indagini condotte sulle aziende agricole sono state avviate alcune ricerche che hanno la finalità di studiare processi di stratificazione che consentono di aumentare la precisione delle stime rispetto a quella ottenibile con l'attuale stratificazione, basata sull'utilizzazione di alcuni caratteri fisici, ad esempio superficie coltivata a frumento. Ampi dettagli riguardanti: gli aspetti metodologici relativi alla formazione dei campioni, il tipo di stimatore, nonché alcune considerazioni sui problemi che sorgono per il fenomeno della mancata risposta totale, sono contenuti in alcuni documenti interni non pubblicati (Russo, 1985 Russo e Falorsi, 1988a; 1988b).

(3) L'ultima indagine condotta utilizzando il campione delle forze di lavoro è stata quella sulle vacanze e gli sports del 1982.

(4) Una descrizione esauriente di tale stimatore è riportata nei Quaderni di Discussione dell'Istat (Russo e Falorsi, 1985).

(5) In generale gli ammontari di popolazione coinvolti nei procedimenti di stima vengono determinati attraverso interpolazioni o estrapolazioni basate sull'utilizzazione delle informazioni relative alla popolazione residente riferite all'inizio e alla fine di ciascun anno.

(6) A questo riguardo precisiamo che lo schema di campionamento non ha subito mutamenti, nel senso che è sempre del tipo a due stadi con stratificazione delle unità primarie; le modificazioni hanno riguardato fondamentalmente la stratificazione, la determinazione della numerosità campionaria, l'allocazione della stessa negli strati elementari e il procedimento di stima (Falorsi, Coccia, Russo e Botta, 1988).

(7) Le classi di età generalmente sono decennali.

(8) Per classe demografica si intende la generica modalità definita dal concatenamento delle modalità del sesso e di quelle relative all'età.

RIFERIMENTI BIBLIOGRAFICI

ALEXANDER C.H., ROEBUCK M.J. (1986) «*Comparison of alternative methods for household estimation*», U.S. Bureau of the Census, U.S.A.

FALORSI P.D., COCCIA G., RUSSO A. (1988) «*Disegno di campionamento, procedura di stima, calcolo e presentazione degli errori campionari*» in «Indagine sugli sport e sulle vacanze: le vacanze degli italiani nel 1985», Note e relazioni, n. 2, Istat, Roma.

FALORSI P.D., COCCIA G., RUSSO A. e BOTTA M. (1988) «*Disegno di campionamento, procedura di stima, calcolo e presentazione degli errori campionari*» in «Indagine sugli sport e sulle vacanze: le vacanze degli italiani nel 1985», Note e relazioni, n. 3, Istat, Roma.

FALORSI S. (1988) «*An Estimate Procedure for the Straightening in Sample Household Surveys*», 1ª Conferenza dell'International Association for Official Statistics, Roma.

RUSSO A. (1985) «*Indagini sulla struttura delle aziende agricole, sulla coltivazione del granoturco, della vite e dell'olivo e sull'allevamento del bestiame bovino*», Istat.

RUSSO A. (1988) «*A Comparative Analisys of Some Estimators Utilized in the Sample Surveys of the Households, 1ª Conferenza dell'International Association for Official Statistics*», Roma.

RUSSO A. e BOTTA M. (1989) «*Un'analisi comparativa di alcuni stimatori utilizzati nelle indagini Istat (documento interno non pubblicato)*».

RUSSO A. e DI TRAGLIA (1985) «*Disegno di campionamento, calcolo e presentazione degli errori campionari*» in «Indagine sulle strutture ed i comportamenti familiari», Istat, Roma.

RUSSO A. e FALORSI P.D. (1985) «*Rilevazione campionaria delle forze di lavoro: Metodologia del campionamento, calcolo e presentazione degli errori campionari*», Quaderni di Discussione n. 6, Istat, Roma.

RUSSO A. e FALORSI P.D. (1988a) «*Ristrutturazione del sistema di indagini agricole*», Istat.

RUSSO A. e FALORSI P.D. (1988b) «*Indagini sulla coltivazione del frumento, del granoturco, dell'olivo e della vite*», Istat.

RUSSO A. e FALORSI P.D., COCCIA G. e GIOVANI P. (1986) «*Disegno di campionamento, calcolo e presentazione degli errori campionari*», in «Indagine sulla lettura ed altri aspetti del tempo libero, anno 1981», Note e relazioni, n. 2, Istat, Roma.

TECNICHE SPECIALI DI STIMA PER PICCOLI DOMINI TERRITORIALI: CONTRIBUTI METODOLOGICI E PRINCIPALI RISULTATI EMPIRICI

di *Piero Demetrio Falorsi*

1. INTRODUZIONE

In Italia, come in molti altri paesi, dagli anni '70 ad oggi si è manifestato un crescente bisogno di informazioni accurate e tempestive riferibili a piccoli domini territoriali. In particolare le Regioni ed altre realtà istituzionali (quali Province, Camere di commercio, comprensori di Comuni) hanno fatto numerose richieste all'Istat per ottenere informazioni attendibili sui fenomeni dell'occupazione e della disoccupazione per livelli territoriali sub-regionali.

Tali richieste hanno fortemente influenzato l'organizzazione dell'indagine Istat sulle forze di lavoro, che è la principale fonte statistica sui fenomeni in oggetto.

In effetti la struttura e la numerosità del campione originale dell'indagine sulle forze di lavoro sono state studiate per ottenere stime affidabili a livello nazionale e regionale per i principali gruppi di popolazione; pertanto, in base al campione originale, risulta problematica l'analisi sia per domini territoriali sub-regionali, sia per le sottoclassi di popolazione di piccola dimensione. La soluzione data dall'Istat ai problemi posti dalla crescente domanda informativa è stata, analogamente a quanto fatto in altri paesi, quella di ampliare il campione sia per quanto riguarda i comuni che per quanto riguarda le famiglie, senza peraltro modificare la strategia campionaria adottata; nel senso che non sono state introdotte modifiche né nel disegno di campionamento, né nello stimatore utilizzato.

I dati che si riportano nel prospetto seguente danno un'idea della dimensione del fenomeno del sovracampionamento negli ultimi dieci anni.

Dimensione del campione Istat sulle forze di lavoro negli anni '73-'88

ANNI	Comuni campione	Famiglie campione
1973-1976 (campione originale)	1.400	83.000
1977-1978	1.500	85.000
1979	1.600	90.000
1980	1.850	114.000
1981-1987	1.900	124.000
1987-1988	2.050	141.000

Nel lavoro di Fabbris, Russo e Sanetti (1988) vengono descritte la storia e le metodologie adottate per i sovracampionamenti dell'indagine sulle forze di lavoro.

L'esigenza di disporre da parte degli enti locali di stime affidabili a livello sub-regionale si è manifestata anche con riferimento ad altre indagini. Ad esempio, la regione Emilia-Romagna ha richiesto, con riferimento alle indagini sulle aziende agricole, la fornitura di stime a livello provinciale, mentre tali indagini sono predisposte per fornire dati a livello regionale. Una richiesta analoga è stata effettuata dalla regione Veneto per quanto riguarda l'indagine Multiscopo.

Per affrontare queste esigenze l'Istat si sta muovendo in un'ottica diversa rispetto a quella adottata nel passato in quanto la soluzione basata soltanto sul sovracampionamento è costosa ed appesantisce la struttura organizzativa ed operativa dell'indagine, potendo inoltre comportare una minore accuratezza nelle operazioni connesse alla rilevazione, con conseguente aumento dell'errore non campionario.

Per le suddette ragioni è diventato imperativo studiare ed iniziare a sperimentare metodologie non troppo costose e facilmente realizzabili sul piano operativo che consentano di migliorare l'affidabilità delle stime a livello di piccolo dominio.

La scelta di un particolare metodo di stima per piccoli domini dipende, come osservano Purcell e Kish (1979), fondamentalmente dal tipo di informazioni disponibili. Una caratteristica comune a tali problemi è costituita dall'esiguità, ed in qualche caso dalla mancanza totale di informazioni a livello di piccolo dominio, mentre si dispone di informazioni a livello aggregato.

La ricerca sviluppata in ambito Istat su tale argomento si è concentrata su due filoni principali:

a) il primo è costituito dai metodi di stima sintetici che, utilizzando le informazioni disponibili a livello aggregato, determinano per il piccolo dominio stime che presentano varianze relativamente basse e livelli di distorsione più elevati di quelli di altri stimatori (Coccia, 1987; Russo e Falorsi 1987; Falorsi 1988; Russo e Falorsi 1988; Russo e Falorsi 1989);

b) il secondo, applicabile al caso di indagini ripetute e periodiche, utilizza un approccio inferenziale basato sui modelli di superpopolazione ed arriva a definire un predittore del parametro oggetto di indagine, relativo alla piccola area, mediante un modello di serie storiche che utilizza le stime ai tempi precedenti (Di Traglia e Russo, 1987; Di Traglia e Falorsi, 1987).

2. STIMATORI SINTETICI

2.1. Premessa

La stima sintetica relativa ad un piccolo dominio territoriale viene determinata mediante una procedura composta essenzialmente di due passi:

— dapprima, relativamente ad un livello territoriale aggregato, vengono determinate le stime della caratteristica di interesse per le differenti sottoclassi in cui è suddivisa la popolazione;

— successivamente tali stime vengono riproporzionate in relazione all'incidenza della sottoclasse all'interno del piccolo dominio.

Come sarà meglio chiarito in seguito, tali stime risultano approssimativamente corrette se la composizione per sottoclasse del piccolo dominio è conosciuta in modo accurato, e se la suddivisione per sottoclassi spiega in modo sufficiente il manifestarsi del fenomeno oggetto di indagine.

La stima sintetica è comunque un modo semplice e poco costoso per ottenere stime relative a piccoli domini territoriali, e ciò spiega il grande successo applicativo del metodo.

Nelle indagini concrete ed in letteratura (vedi ad esempio: Purcell e Linacre, 1976; Gonzales e Waksberg, 1973; Gonzales ed Hoza, 1978) sono state proposte differenti forme di stimatori sintetici, corrispondenti a modi diversi di combinare le informazioni campiona-

rie con quelle desumibili da altre fonti (registri anagrafici, dati censuari, ecc.).

La ricerca sull'argomento in oggetto, sviluppata in ambito Istat, ha avuto il principale obiettivo di approfondire la teoria delle stime sintetiche in modo da renderla applicabile alla realtà delle strategie campionarie condotte dall'Istat.

Nel lavoro di Coccia (1987), presentato al Convegno SIS di Perugia, viene proposto uno stimatore sintetico da adottare nelle indagini sulle aziende agricole basate su un disegno di campionamento ad uno stadio stratificato. Nel lavoro suddetto viene, inoltre, derivata l'espressione dell'errore quadratico medio e si suggerisce un metodo per la valutazione della distorsione.

Nel lavoro di Falorsi e Russo (1987), presentato al Convegno SIS di Perugia, è stato studiato a livello teorico un primo tipo di stimatore sintetico da utilizzare nelle indagini Istat sulle famiglie che, come è noto, si fondano su di una strategia campionaria basata su di un disegno a due stadi con stratificazione delle unità primarie ed uno stimatore del rapporto post-stratificato.

Nella comunicazione in questione vengono fornite le espressioni della distorsione e della varianza dello stimatore proposto. Si espongono, inoltre, due criteri per la valutazione dell'efficienza del suddetto stimatore rispetto allo stimatore del rapporto post-stratificato. Il primo criterio calcola la distorsione dello stimatore sintetico in base ad informazioni non campionarie; mentre il secondo stima la distorsione utilizzando i dati campionari mediante un modello di superpopolazione.

Nella comunicazione di Falorsi (1988), presentata al Convegno IAOS di Roma, viene illustrato un esperimento avente il fine di valutare in una situazione concreta le caratteristiche dello stimatore sintetico introdotto nel lavoro di Falorsi e Russo (1987). L'esperimento è stato condotto sulla base dei dati censuari e relativamente al disegno di campionamento dell'indagine Istat sulle forze di lavoro.

Il lavoro, in cui vengono esaminati in maniera più approfondita gli aspetti teorici ed empirici connessi all'utilizzazione degli simatori sintetici, è quello di Russo e Falorsi (1988), presentato al Convegno FO.LA. di Bressanone.

Nella comunicazione in oggetto, con riferimento all'indagine sulle forze di lavoro e relativamente al caso di domini sub-regionali co

stituiti da parti di strati elementari, vengono studiati, valutati e messi a confronto alcuni stimatori per piccole aree. Oltre allo stimatore sintetico vengono esaminati lo stimatore del rapporto post-stratificato ed il cosiddetto stimatore composto, espresso come un opportuna media ponderata degli stimatori post-stratificato e sintetico. La valutazione della distorsione, della varianza e dell'errore quadratico medio di ciascuno stimatore, nonché dell'efficienza degli stimatori sintetico e composto rispetto allo stimatore post-stratificato è effettuata mediante uno studio empirico basato sull'utilizzazione di dati ed informazioni relative al censimento della popolazione del 1981.

Dalle risultanze sperimentali del lavoro in questione emergono indicazioni molto confortanti sull'uso dello stimatore sintetico che presenta errori quadratici medi molto bassi e distorsioni relative accettabili.

Nel prossimo paragrafo verranno illustrati i principali risultati del lavoro suddetto.

2.2. Descrizione dei principali risultati

2.2.1 Simbologia

Con riferimento al disegno campionario attualmente adottato per l'indagine sulle forze di lavoro e relativamente ad una generica provincia, indichiamo con:

- h indice di strato
- H insieme degli strati definiti in una provincia
- i indice di comune
- j indice di famiglia
- a indice di classe demografica [1]
- A insieme delle classi demografiche
- N_h insieme dei comuni compresi nello strato h
- n_h insieme dei comuni-campione nello strato h (per l'indagine in esame $n_h = 1$ per ogni $h \in H$)
- M_{hi} insieme delle famiglie residenti nel comune (h i)
- m_{hi} insieme delle famiglie-campione nel comune (h i)
- P_{hi} popolazione residente nel comune (h i)
- P_h popolazione residente nello strato h

$_aP_{hij}$ numero di componenti nella classe demografica a appartenenti alla famiglia hij

$_aP_h$ popolazione residente nello strato h ed appartenente alla classe demografica a

$_aP$ popolazione residente nella provincia ed appartenente alla classe demografica a

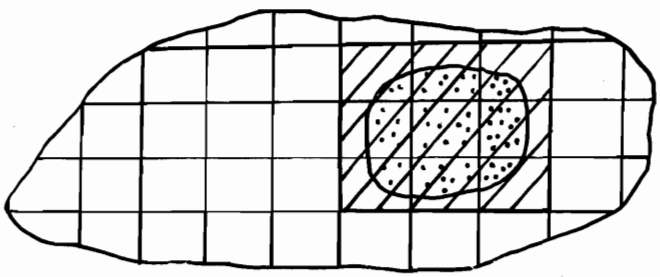
K_{1hi} peso base attribuito al comune (h i)

$K_{hi} = K_{1hi}\left(\frac{M_{hi}}{m_{hi}}\right)$ peso attribuito alla generica famiglia nel comune (h i)

2.2.2 *Parametri oggetto di stima*

Consideriamo il dominio sub-provinciale «d» e supponiamo che esso sia incluso in un sottoinsieme di strati (che indichiamo con «D») dell'insieme degli H strati definiti nella generica provincia.

La situazione è illustrata nella figura sotto riportata: la parte tratteggiata rappresenta il sottoinsieme D, quella punteggiata indica il piccolo dominio d.



Sia ora $_aY_{hij}$ il totale del generico carattere y oggetto di indagini relativo ai $_aP_{hij}$ elementi della sottoclasse a nella famiglia (hij).

Il totale del generico carattere y riferito alla data provincia è pertanto definito dall'espressione:

$$Y = \sum_{h\epsilon H} \sum_{a\epsilon A} \sum_{i\epsilon N_h} \sum_{j\epsilon M_{hi}} {}_aY_{hij} =$$

$$= \sum_{h\epsilon H} \sum_{a\epsilon A} {}_aY_h = \sum_{h\epsilon H} Y_h = \sum_{a\epsilon A} {}_aY \qquad (1)$$

Con riferimento al dominio d, il totale del generico carattere y è invece espresso dalla relazione:

$$^{d}Y = \sum_{h\epsilon H} \sum_{a\epsilon A} \sum_{i\epsilon {}^{d}N_h} \sum_{j\epsilon M_{hi}} {}_{a}Y_{hij}$$

$$= \sum_{h\epsilon H} \sum_{a\epsilon A} {}^{d}_{a}Y_h = \sum_{h\epsilon H} {}^{d}Y_h = \sum_{a\epsilon A} {}^{d}_{a}Y \qquad (2)$$

in cui $^{d}N_h$ indica l'insieme dei comuni appartenenti congiuntamente allo strato h ed al dominio d.

2.2.3 *Stimatori*

Ricordiamo che nella sperimentazione in oggetto sono stati presi in esame uno stimatore del rapporto con post-stratificazione secondo le modalità combinate del sesso e delle classi di età [2], lo stimatore sintetico e lo stimatore composto.

L'espressione formale dello stimatore del rapporto combinato e post-stratificato è data da:

$$^{d}\hat{Y}_{POS} = \frac{\sum_{h\epsilon H} \sum_{a\epsilon A} {}_{a}\hat{Y}_h \, \delta_h}{{}_{a}\hat{P}} \, {}_{a}P \qquad (3)$$

in cui:

$${}_{a}\hat{Y}_h = \sum_{j\epsilon m_{hi}} K_{hi} \, {}_{a}Y_{hij} \qquad (4)$$

$${}_{a}\hat{P} = \sum_{h\epsilon H} \sum_{j\epsilon m_{hi}} K_{hi} \, {}_{a}P_{hij} \qquad (5)$$

e δ_h è una variabile indicatrice che assume valore 1 se il comune campione selezionato nello strato h appartiene all'insieme $^{d}N_h$, valore 0 altrimenti.

Per quanto concerne gli stimatori sintetici sono state prese in esame tre differenti versioni, indicate rispettivamente con:

$$^{d}\hat{Y}_{SIN/1};\ ^{d}\hat{Y}_{SIN/2};\ ^{d}\hat{Y}_{SIN/3}$$

Il primo stimatore è definito dall'espressione:

$$ {}^{d}\hat{Y}_{SIN/1} = \sum_{a \epsilon A} \frac{\sum_{h \epsilon D} {}_{a}\hat{Y}_{h} \frac{{}^{d}_{a}P_{h}}{{}_{a}P_{h}}}{{}_{a}\hat{P}} {}_{a}P \qquad (6) $$

dove:

${}^{d}_{a}P_{h}$ indica il sottoinsieme di ${}_{a}P_{h}$ appartenente al dominio d.

Lo stimatore ${}^{d}\hat{Y}_{SIN/2}$ è espresso dalla relazione:

$$ {}^{d}\hat{Y}_{SIN/2} = \sum_{a \epsilon A} \frac{\sum_{h \epsilon D} {}_{a}\hat{Y}_{h} \frac{\sum_{h \epsilon D} {}^{d}_{a}P_{h}}{\sum_{h \epsilon D} {}_{a}P_{h}}}{{}_{a}\hat{P}} {}_{a}P \qquad (7) $$

La terza versione di stimatore sintetico, infine, è definita dall'espressione:

$$ {}^{d}\hat{Y}_{SIN/3} = \sum_{a \epsilon A} \sum_{h \epsilon H} {}_{a}\hat{Y}_{h} \frac{\sum_{h \epsilon D} {}^{d}_{a}P_{h}}{{}_{a}P} \qquad (8) $$

La differenza tra gli stimatori sintetici e lo stimatore post-stratificato risulta evidente confrontando la relazione (3) con le relazioni (6), (7) e (8).

Lo stimatore post-stratificato utilizza l'informazione dello strato $h \epsilon D$ solo se il comune campione dello strato h appartiene all'insieme ${}^{d}N_{h}$; nel caso in cui in tutti gli strati del dominio D non venga estratto un comune campione appartenente alla piccola area d, diviene addirittura impossibile costruire la stima relativa al piccolo dominio.

Nelle espressioni degli stimatori sintetici, invece, vengono coinvolti la stima ${}_{a}Y_{h}$, la cui determinazione si basa su tutte le informazioni appartenenti allo strato h, e dei fattori specifici di aggiustamento per ciascuna classe demografica.

I tre stimatori sintetici differiscono fra loro per il livello territoriale rispetto al quale viene costruita la correzione sintetica.

Relativamente alla sottoclasse a, in SIN/1 la correzione sintetica viene applicata sulla stima del singolo strato h; in SIN/2 la corre-

zione viene applicata sulla stima relativa al dominio D ed in SIN/3 sulla stima relativa alla intera provincia.

Lo stimatore sintetico si basa sull'ipotesi che le medie di classe demografica a livello aggregato siano uguali, o molto vicine, a quelle relative al piccolo dominio; più precisamente, con riferimento ad esempio allo stimatore SIN/1, l'ipotesi consiste nel supporre che, per ogni classe demografica e per ogni strato, il rapporto a livello aggregato ${}_aY_h/{}_aP_h$, sia uguale, o molto vicino, a quello espresso da ${}^d_aY_h/{}^d_aP_h$.

Proprio per questo lo stimatore sintetico presenta vari problemi; in generale esso è distorto in quanto le medie a livello aggregato possono essere differenti da quelle a livello di piccolo dominio. Le differenze derivano essenzialmente da due motivi: da un lato la suddivisione in classi demografiche può non spiegare in modo sufficientemente adeguato la variabilità delle medie parziali; dall'altro gli stimatori in questione non consentono di tener conto di fattori locali che potrebbero influire sulle medie a livello di classi demografiche.

Nella ricerca, come abbiamo già detto, viene studiato anche lo stimatore composto espresso come media ponderata dello stimatore sintetico e dello stimatore del rapporto post-stratificato.

La particolare convenienza dell'impiego di tale stimatore è connessa al fatto che esso gode della proprietà di mediare le situazioni estreme sia in termini di distorsione, sia in termini di varianza campionaria. Pertanto in un'assegnata situazione concreta, tale stimatore può risultare più vantaggioso, rispetto alle sue due componenti separatamente considerate. Da un lato infatti lo stimatore sintetico presenta una varianza piccola, ma può essere affetto da una distorsione molto forte se, come abbiamo precedentemente osservato, l'assunzione di omogeneità non è soddisfatta. Dall'altro lo stimatore post-stratificato risulta asintoticamente corretto, ma può presentare una varianza molto alta se la dimensione del campione, costituito dalle unità primarie che cadono nel dominio d, è molto piccola.

Una valutazione empirica circa il comportamento e la precisione dello stimatore composto rispetto agli stimatori diretto e sintetico, può trovarsi negli studi di Gonzales e Waksberg (1975) e di Schaible, Brock e Schnaick (1977).

Tenendo presente la precedente trattazione sullo stimatore sintetico, è possibile definire le seguenti tre forme di stimatore composto:

$$ {}^{d}\hat{Y}_{COM/1} = \alpha_1 \, {}^{d}\hat{Y}_{POS} + (1 - \alpha_1) \, {}^{d}\hat{Y}_{SIN/1} \tag{9} $$

$$ {}^{d}\hat{Y}_{COM/2} = \alpha_2 \, {}^{d}\hat{Y}_{POS} + (1 - \alpha_2) \, {}^{d}\hat{Y}_{SIN/2} \tag{10} $$

$$ {}^{d}\hat{Y}_{COM/3} = \alpha_3 \, {}^{d}\hat{Y}_{POS} + (1 - \alpha_3) \, {}^{d}\hat{Y}_{SIN/3} \tag{11} $$

in cui α_1, α_2 e α_3 sono pesi scelti in modo da minimizzare l'errore quadratico medio dei rispettivi stimatori composti [3].

2.2.4 *Varianze e distorsioni degli stimatori proposti*

Il problema di determinare le varianze e le distorsioni degli stimatori precedentemente esposti, che presentano la caratteristica di essere non lineari, è stato affrontato nell'ottica di cercare una soluzione approssimata. Il metodo utilizzato, per raggiungere tale scopo, approssima la varianza e la distorsione degli stimatori considerati mediante la varianza e la distorsione dei termini di ordine lineare dello sviluppo in serie di Taylor degli stimatori medesimi.

Pertanto, tenendo presenti le approssimazioni introdotte ed il disegno campionario adottato nell'indagine sulle forze di lavoro, nello studio di Russo e Falorsi (1988) sono derivate le espressioni della varianza e della distorsione degli stimatori sopra descritti.

In questa nota illustriamo tali espressioni limitatamente allo stimatore SIN/1:

$$ V\left({}^{d}\hat{Y}_{SIN/1}\right) = \sum_{h \in H} \sum_{i \in N_h} K_{1hi}^{-1} \left(Z_{hi} \, K_{1hi} - Z_h\right)^2 + $$

$$ + \sum_{i \in N_h} K_{hi} \left(M_{hi} - m_{hi}\right) S_{hi}^2 \tag{12} $$

$$ B\left({}^{d}\hat{Y}_{SIN/1}\right) = \sum_{h \in H} \sum_{a \in A} \left({}_{a}Y_h \frac{{}^{d}_{a}P_h}{{}_{a}P_h} - {}^{d}_{a}Y_h\right) \tag{13} $$

in cui:

$$ Z_h = \sum_{i \in N_h} Z_{hi} = \sum_{i \in M_h} \sum_{j \in M_{hi}} Z_{hij} \tag{14} $$

$$Z_{hij} = \sum_{i \epsilon N_h} \sum_{j \epsilon M_{hi}} \sum_{a \epsilon A} \left({}_aY_{hij} \frac{{}^dP_h}{{}_aP_h} - {}^dR_a \, {}_aP_{hij} \right) \tag{15}$$

$${}^dR_a = \frac{\sum_{h \epsilon D} {}_aY_h \frac{{}^dP_h}{{}_aP_h}}{{}_aP} \tag{16}$$

$$S^2_{hi} = (M_{hi} - 1)^{-1} \sum_{j \epsilon M_{hi}} (Z_{hij} - Z_{hi} \, M_{hi}^{-1})^2 \tag{17}$$

L'errore quadratico medio di $({}^d\hat{Y}_{SIN/1})$ è, come noto, espresso dalla relazione:

$$MSE \, ({}^d\hat{Y}_{SIN/1}) = V \, ({}^d\hat{Y}_{SIN/1}) + B^2 \, ({}^d\hat{Y}_{SIN/1}) \tag{18}$$

2.2.5 *Analisi dei risultati*

Gli aspetti essenziali dello studio empirico condotto con riferimento alle 84 USL della regione Lombardia sono:

a) ai fini della formazione del disegno campionario e della determinazione della distorsione e della varianza di ciascuno degli stimatori considerati, si sono utilizzati dati ed informazioni tratte dal XII Censimento generale della popolazione (ottobre 1981);

b) tutti i comuni (1546) della regione in questione sono stati suddivisi in strati, attraverso un procedimento di stratificazione uguale a quello attualmente adottato nell'indagine sulle forze di lavoro;

c) avendo scelto l'USL come piccolo dominio di interesse, si è proceduto anche all'attribuzione dei suddetti 1546 comuni alle 84 USL costituite nella regione;

d) per ciascuna USL, con riferimento alle due stime:

— numero di occupati
— numero di disoccupati

sono state calcolate la distorsione relativa percentuale rispetto al valore vero, l'errore relativo percentuale, la distorsione relativa percentuale rispetto all'errore quadratico medio e l'efficienza relativa.

Le considerazioni svolte nelle pagine seguenti sono basate sui valori medi provinciali delle quantità descritte al punto d).

Per lo stimatore SIN/1, tali valori medi sono espressi dalle seguenti relazioni:

$$- B_r\left({}^{d}\hat{Y}_{SIN/1}\right) = d_p^{-1} \sum_{d \epsilon d_p} \frac{B\left({}^{d}\hat{Y}_{SIN/1}\right)}{{}^{d}\hat{Y}} 100 \tag{19}$$

per la distorsione relativa percentuale media;

$$- \epsilon\left({}^{d}\hat{Y}_{SIN/1}\right) = d_p^{-1} \sum_{d \epsilon d_p} \frac{\sqrt{MSE\left({}^{d}\hat{Y}_{SIN/1}\right)}}{{}^{d}\hat{Y}} 100 \tag{20}$$

per distorsione relativa percentuale media rispetto all'errore quadratico medio;

$$- \epsilon\left({}^{d}\hat{Y}_{SIN/1}\right) = d_p^{-1} \sum_{d \epsilon d_p} \sqrt{\frac{B^2\left({}^{d}\hat{Y}_{SIN/1}\right)}{MSE\left({}^{d}\hat{Y}_{SIN/1}\right)}} 100 \tag{21}$$

per l'errore relativo percentuale medio;

$$- Eff\left({}^{d}\hat{Y}_{SIN/1}\right) = d_p^{-1} \sum_{d \epsilon d_p} \frac{V\left({}^{d}\hat{Y}_{POS}\right)}{MSE\left({}^{d}\hat{Y}_{SIN/1}\right)} \tag{22}$$

per l'efficienza relativa media, rispetto allo stimatore post-stratificato; in cui d_p indica l'insieme di USL nella generica provincia.

L'esame della tabella 1 consente di effettuare una prima valutazione degli stimatori proposti in termini di distorsione relativa percentuale. Riteniamo opportuno sottolineare che in essa non figura lo stimatore post-stratificato in quanto, per l'uso dell'approssimazione di Taylor limitata ai termini di ordine lineare, la distorsione di tale stimatore risulta nulla; tuttavia, come dimostrano alcune esperienze effettuate in altri paesi, la distorsione degli stimatori post-stratificati nelle indagini condotte su larga scala presenta valori generalmente trascurabili ed, in ogni caso, di entità più piccola rispetto alle distorsioni degli stimatori sintetici.

Studiando in dettaglio i valori riportati nella tabella in questione, notiamo un forte aumento della distorsione quando si passa da-

gli occupati ai disoccupati. Per gli occupati, infatti, tutti gli stimatori mostrano distorsioni relative percentuali trascurabili (< 2,5%), i cui livelli più bassi si registrano per gli stimatori SIN/1 e COM/1.

Per i disoccupati, limitando dapprima l'osservazione ai tre stimatori sintetici, vediamo che i livelli di distorsione dello stimatore SIN/1 sono sistematicamente più piccoli, con un campo di variazione: (0,55) — (10,98); estendendo poi l'esame anche agli stimatori composti si osserva che lo stimatore COM/1 fa registrare, in assoluto, i livelli più bassi di distorsione, con un campo di variazione: (0,29) — (7,16).

Nella tabella 2 sono esposti i valori dell'errore relativo percentuale di tutti gli stimatori considerati nel presente lavoro. Sulla base di tali valori, si evince che lo stimatore SIN/3 fornisce i risultati migliori in tutte le province, sia per gli occupati che per i disoccupati; lo stimatore POS, secondo le attese, è quello caratterizzato dai valori più alti.

Passando ora all'esame della distorsione percentuale sull'errore quadratico medio, dall'osservazione dei dati raccolti nella tabella 3 emerge che fra gli stimatori sintetici lo stimatore SIN/1 è quello a cui corrispondono i valori più bassi della distorsione. Se si estende l'esame anche agli stimatori composti vediamo che lo stimatore COM/1, espresso come media ponderata degli stimatori SIN/1 e POS, fa registrare una decisa diminuzione dei valori della distorsione rispetto a quelli dello stimatore SIN/1.

Nella tabella 4, infine, sono indicate le misure dell'efficienza relativa degli stimatori esaminati rispetto allo stimatore post-stratificato. Esse mostrano che i valori più elevati si registrano per lo stimatore SIN/3 e per lo stimatore composto espresso come combinazione del suddetto stimatore e di quello post-stratificato.

In conclusione riteniamo opportuno osservare che gli stimatori sintetici, analizzati nello studio in questione, risultano di gran lunga preferibili allo stimatore post-stratificato. Un giudizio globale sugli stimatori considerati, desunto attraverso la considerazione congiunta degli aspetti riguardanti la distorsione, l'errore quadratico medio, l'efficienza e la complessità computazionali, sembra condurre alla conclusione che lo stimatore migliore sia SIN/3.

Il giudizio è basato essenzialmente sulle seguenti considerazioni:

— SIN/3 presenta un errore quadratico medio notevolmente inferiore a quello degli altri stimatori;

— la distorsione di SIN/3 è più elevata di quella degli altri stimatori. Essa si mantiene, comunque, sempre a livelli accettabili e, tali da far rimanere l'errore quadratico medio di SIN/3 inferiore agli altri;

— nello studio in oggetto le correzioni sintetiche sono state determinate in base alle risultanze censuarie. Se venisse accettato il principio di utilizzare la stima sintetica come metodo standard per l'ottenimento delle stime per le piccole aree, sarebbe preferibile — data la non aggiornabilità dei dati censuari — determinare i fattori di correzione sintetici mediante l'utilizzo di opportuni modelli demografici. Ora, come è noto, l'attendibilità dei dati risultanti da tali modelli aumenta al crescere del livello territoriale di riferimento. Anche per tale ragione diviene preferibile lo stimatore SIN/3 in cui la correzione sintetica è applicata sulla stima relativa all'intera provincia;

— SIN/3 è lo stimatore di più facile determinazione, in quanto richiede la costruzione di un numero minimo di fattori correttivi.

Tabella 1 - Distorsione relativa percentuale: valori medi provinciali

PROVINCIA	OCCUPATI						DISOCCUPATI					
	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3
Varese	0,87	1,61	1,87	0,84	1,56	1,85	7,45	11,11	10,27	5,51	8,13	8,74
Como	1,18	1,78	2,00	1,13	1,72	1,98	7,19	9,23	20,87	5,04	6,98	17,69
Sondrio	0,84	0,52	0,87	0,78	0,49	0,87	10,98	15,24	21,75	7,16	9,18	14,45
Milano	0,37	0,73	2,21	0,36	0,70	2,18	6,24	9,21	24,66	4,66	6,82	18,4
Bergamo	0,75	1,42	1,99	0,69	1,31	1,94	7,17	7,78	16,60	5,20	5,74	14,03
Brescia	0,41	0,58	2,77	0,37	0,51	2,49	2,97	4,12	13,57	2,07	2,96	10,49
Pavia	0,10	0,44	2,86	0,07	0,33	2,01	0,55	0,74	15,17	0,29	0,41	8,17
Cremona	1,05	0,96	1,76	1,02	0,93	1,73	2,17	21,37	18,56	1,38	1,51	13,11
Mantova	0,33	1,16	1,97	0,32	1,11	1,94	2,54	4,83	8,22	1,86	3,59	7,06

Tabella 2 - Errore relativo percentuale: valori medi provinciali

PROVINCIA	OCCUPATI							DISOCCUPATI					
	POS	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3	POS	SIN/1	SIN/2	SIN/3	COM/1	COM/2
Varese	40,7	7,3	6,9	2,4	7,1	6,7	2,4	60,1	19,8	18,8	10,4	29,9	29,1
Como	49,9	9,8	9,2	3,4	9,6	9,0	3,4	83,1	30,3	28,1	14,1	43,2	39,9
Sondrio	38,0	10,0	9,3	2,2	9,6	8,9	2,2	54,0	31,3	35,1	28,8	29,0	31,6
Milano	40,4	6,9	7,5	2,9	6,7	7,2	2,9	66,5	15,6	14,6	11,6	33,5	32,0
Bergamo	40,3	10,7	9,7	4,2	10,3	9,3	4,2	64,3	16,8	15,2	9,5	33,5	30,7
Brescia	34,4	10,5	9,8	4,5	9,8	9,3	4,3	57,7	29,4	26,6	15,6	31,8	29,5
Pavia	6,6	4,3	4,4	3,9	3,5	3,6	3,4	25,5	11,2	10,7	10,1	17,7	17,3
Cremona	23,6	4,5	4,1	3,2	4,5	4,1	3,2	42,9	13,5	12,0	11,6	25,7	23,7
Mantova	37,6	7,4	7,1	4,1	7,3	6,9	4,1	62,4	16,4	14,8	8,6	31,8	29,2

Tabella 3 - Percentuale della distorsione sull'errore quadratico medio: valori medi percentuali

PROVINCIA	OCCUPATI						DISOCCUPATI					
	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3
Varese	2,60	6,63	46,29	2,52	6,44	45,83	5,15	11,67	31,25	3,81	8,48	26,83
Como	2,15	5,76	31,61	2,09	5,55	31,35	2,94	8,06	60,05	2,03	6,04	50,67
Sondrio	0,92	0,48	20,17	0,85	0,45	20,12	12,14	16,78	43,78	8,35	10,15	28,88
Milano	0,40	1,76	52,01	0,39	1,70	51,44	3,54	8,43	56,44	2,68	6,14	41,78
Bergamo	0,69	3,24	24,39	0,65	2,97	23,75	3,89	6,60	48,30	2,83	4,74	41,03
Brescia	0,24	0,96	34,94	0,22	0,84	31,87	0,74	2,05	42,92	0,51	1,44	33,73
Pavia	0,08	1,86	51,26	0,06	1,41	35,04	0,09	0,13	47,48	0,05	0,07	25,25
Cremona	5,35	5,89	29,19	5,16	5,73	28,72	0,51	2,07	39,07	0,32	1,29	27,61
Mantova	0,26	3,27	24,32	0,25	3,17	23,92	0,85	3,23	17,74	0,62	2,51	15,19

Tabella 4 - Efficienza degli stimatori rispetto allo stimatore post-stratificato: valore medi provinciali

PROVINCIA	OCCUPATI						DISOCCUPATI					
	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3	SIN/1	SIN/2	SIN/3	COM/1	COM/2	COM/3
Varese	5,98	6,06	22,45	6,07	6,15	22,49	1,72	1,80	3,54	2,00	2,06	3,70
Como	5,12	5,46	14,80	5,23	5,56	14,84	1,57	1,79	3,23	1,87	2,06	3,41
Sondrio	4,08	4,38	17,52	4,21	4,50	17,55	1,64	1,41	1,95	1,94	1,74	2,22
Milano	5,58	5,24	16,25	5,69	5,36	16,30	1,61	1,72	2,39	1,91	2,01	2,62
Bergamo	3,68	3,95	10,04	3,82	4,08	10,11	1,58	1,75	2,95	1,88	2,03	3,14
Brescia	3,26	3,47	8,75	3,43	3,63	8,84	1,46	1,62	3,31	1,78	1,91	3,49
Pavia	1,59	1,53	1,76	1,89	1,83	2,02	1,04	1,09	1,22	1,44	1,48	1,59
Cremona	5,38	5,62	7,17	5,47	5,71	7,24	1,30	1,45	1,53	1,64	1,76	1,83
Mantova	5,19	5,46	9,31	5,29	5,56	9,37	1,68	1,91	3,21	1,96	2,17	3,38

3. CONSIDERAZIONI CONCLUSIVE E FUTURI ITINERARI DI RICERCA

Nelle considerazioni svolte nell'introduzione abbiamo sottolineato che, nel contesto delle indagini Istat per le quali il dominio territoriale di studio è la regione geografica, il problema di ottenere stime affidabili per domini territoriali subregionali (province, aree funzionali, bacini del lavoro) è stato affrontato unicamente attraverso un'aumento della dimensione campionaria.

D'altra parte, le ricerche da noi condotte sul terreno dei metodi di stima per piccoli domini, i cui principali risultati teorici ed empirici sono stati riassunti nella presente nota, hanno mostrato che l'utilizzazione di metodi di stima speciali per piccole aree consente di migliorare il livello di precisione delle stime rispetto a quello ottenibile mediante i metodi di stima attualmente adottati.

Tale risultato costituisce un utile suggerimento al fine di dare soluzione al problema in esame con un'ottica diversa rispetto a quella basata soltanto sul sovracampionamento.

Riteniamo, cioè, che in un prossimo futuro le stime relative a domini sub-regionali potranno essere ottenute attraverso procedure basate su stimatori per piccole aree. Tuttavia, nei casi in cui l'utilizzazione di tali procedure non dovesse garantire il livello atteso di precisione, si farà ricorso ad una soluzione basata anche sul sovracampionamento.

Ciò detto, descriviamo ora per sommi capi, quali sono i futuri itinerari della ricerca Istat sul terreno delle stime per piccole aree.

Ricordiamo, come già illustrato nell'introduzione, che la linea di ricerca Istat sull'argomento in oggetto si è sviluppata essenzialmente secondo due approcci: il primo fondato sui metodi di stima sintetici; il secondo basato sui modelli di superpopolazione.

Per quanto riguarda il primo approccio, riteniamo che sia necessario studiare e valutare sperimentalmente, ponendoli a confronto con gli stimatori sintetici, altri tipi di stimatori come, ad esempio, gli stimatori di regressione generalizzata (Gonzales ed Hoza, 1978).

Riteniamo inoltre che la metodologia sinora usata per la valutazione dei metodi di stima per piccole aree debba essere ulteriormente affinata, in special modo per quanto concerne i problemi relativi alla valutazione dell'errore quadratico medio.

Nell'ambito del primo approccio, l'Istat sta approfondendo una metodologia alternativa, che determina le caratteristiche distributive degli stimatori in base a simulazioni con il metodo di Montecarlo. Il metodo della simulazione consentirebbe da un lato di ottenere ulteriori conferme dei risultati sinora raggiunti; dall'altro permetterebbe di superare alcuni limiti del metodo precedentemente utilizzato come, per esempio, il fatto di considerare nulla la distorsione dello stimatore post-stratificato.

Per quanto riguarda il secondo approccio, ricordiamo che esso è applicabile nel caso di indagini ripetute e periodiche. Per tali indagini è possibile arrivare a definire un predittore del parametro oggetto di indagine relativo alla piccola area mediante un modello di serie storiche che determina la stima in un determinato istante in tempo in funzione delle stime relative ai tempi precedenti.

Tale approccio è stato sviluppato unicamente a livello teorico; mentre mancano a tutt'oggi evidenze empiriche sulla praticabilità ed affidabilità dei metodi proposti. È essenziale, quindi, sviluppare ricerche empiriche in cui le caratteristiche delle stime per piccole aree, ottenute in base al modello di serie storiche, vengano confrontate con quelle ottenibili in base ad altri metodi.

NOTE

(1) Le classi sono formate in base alle combinazioni delle modalità del sesso e delle seguenti classi di età: 14-19; 20-29; 30-54; 55-59, 60 ed oltre.

(2) Vedi nota 1.

(3) Come si dimostra in Falorsi e Russo (1988) il valore ottimale di α_h (h = 1,2,3) è funzione dei valori MSE, V e B e r sia di $^{d}\hat{Y}_{POS}$ che di $^{d}\hat{Y}_{SIN/h}$. Nello studio suddetto sono stati determinati i valori di MSE, V e B sia di $^{d}\hat{Y}_{POS}$ che di $^{d}\hat{Y}_{SIN/h}$, ma non disponendo dei valori di r, si è posto: r = 0; 0,1; 0,2; In tal modo sono stati determinati tre valori per ciascun peso α_h (h = 1,2,3). Conseguentemente nel lavoro testé citato, sono stati analizzati nove stimatori composti; nella presente pubblicazione tuttavia vengono presentati i risultati sperimentali degli stimatori composti ricavati sotto l'ipotesi che il coefficiente di correlazione sia uguale a 0.

(4) Nel lavoro di Russo e Falorsi (88) la distorsione dello stimatore $^{d}\hat{Y}_{POS}$ è stata posta uguale a zero. Per ottenere una distorsione non nulla si sarebbe dovuto considerare lo sviluppo in serie di Taylor comprendente almeno i termini di ordine quadratico, che avrebbero comportato ulteriori complessità di calcolo. Tuttavia, come mostrano alcuni studi condotti all'estero, le stime di indagini concrete, ottenute mediante l'impiego di stimatori post-stratificati, sono generalmente affette da livelli di distorsione trascurabili.

RIFERIMENTI BIBLIOGRAFICI

COCCIA G. (1987), «*Stima per piccole aree: problemi e valutazioni relativi a dati di carattere economico*», Atti del Convegno SIS «Informazione ed analisi statistica per aree regionali e subregionali», Perugia 5-6 ottobre 1987.

DI TRAGLIA M., FALORSI S. (1987), «*Approccio model-based nel campionamento in tempi successivi per piccole aree*», Atti del Convegno SIS «Informazione ed analisi statistica per aree regionali e sub regionali», Perugia 5-6 ottobre 1987.

DI TRAGLIA M., RUSSO A. (1988) «*A methodology to obtain preliminary estimates from repeated surveys*», Atti del Seminario di Metodologia Statistica, Conferenza degli statistici Europei, Ginevra, 1988.

FABBRIS L., RUSSO A., SANETTI I. (1987), «*Storia e proposte in tema di campionamento a livello regionale, provinciale e sub-provinciale per indagini sulle forze di lavoro*». Quaderno FOLA n. 4.

FALORSI P.D., RUSSO A. (1987), «*Un metodo di stima sintetica per piccoli domini territoriali nelle indagini Istat sulle famiglie*», Atti del Convegno SIS «Informazione ed analisi statistica per aree regionali e sub-regionali», Perugia 5-6 ottobre 1987.

FALORSI P.D. (1988), «*Small Area Estimation of the Labour Force Sample Survey: an Empirical Comparison of two Estimators*», Atti dell'International Association Survey Statistican, Roma.

GONZALES M.E., WAKSBERG J. (1973), «*Estimation of the error of synthetic estimates*», First meeting of the International Association of Survey Statisticians.

GONZALES M.E., HOZA C. (1978), «*Small area estimation with application to unemployement and housing estimates*», Journal of the America Statistical Association 73, p. 7-15.

PURCELL N.J., LINACRE S. (1976), «*Techniques for the estimation of small area characteristics*», Australian Statistical Conference, Melbourne.

PURCELL N.J. e KISH L. (1979), «*Estimation for small domains, Biometrics*», p. 365.

RUSSO A., FALORSI P.D. (1988), «*Valutazione di alcune tecniche di stima per piccole aree per l'indagine sulle forze di lavoro*». «Atti del Seminario: Forze di lavoro: Disegno dell'indagine e analisi strutturali», Bressanone (BZ)» 14-16 settembre 1988.

RUSSO A., FALORSI P.D. (1989), «*Un analisi comparativa di alcuni metodi di stima per piccole aree*», Comunicazione al convegno «Analisi statistica di dati territoriali, metodi, tecnologie, applicazioni, Bari 16-17 Marzo 1989.

SCHAIBLE W.L., BROCK D.B. E SCNAICK G.A. (1977), «*An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics*», Proceedings of the American Statistical Association, Social Statistical, 1017-1021.

WOODRUFF R.S. (1971), «*Simple method for approximating variance of complicated estimate*», Journal of the American Statistical Association, 411-414.

CONSIDERAZIONI METODOLOGICHE SULL'USO DI INFORMAZIONI AUSILIARIE NELLE INDAGINI CAMPIONARIE ISTAT

di *Mario Di Traglia*

1. INTRODUZIONE

L'uso delle informazioni disponibili sulla popolazione oggetto di studio occupa, nella letteratura sulle indagini campionarie, un posto rilevante.

Tale importanza è dovuta all'utilità di queste informazioni, sia nella fase di progettazione del disegno di campionamento che in quella di costruzione delle procedure di stima.

Viene intesa, in questa sede, con il termine informazione ausiliaria, l'insieme delle modalità assunte da alcune variabili (oggetto di indagine e non) su tutte o parte delle unità della popolazione, utili al fine della predisposizione del disegno di campionamento e/o dei procedimenti di stima.

È possibile individuare, nell'ambito della teoria del campionamento due situazioni estreme nell'uso di informazioni disponibili. Una è rappresentata dalla strategia campionaria costituita dal disegno di campionamento semplice e dallo stimatore «diretto», l'altra è rappresentata dal disegno di campionamento ragionato e dall'uso di stimatori del rapporto o di regressione. Nella prima non vengono utilizzate informazioni ausiliarie, nella seconda tali informazioni condizionano pesantemente sia le scelte delle unità da rilevare che le procedure di stima.

È possibile notare che, nella teoria classica del campionamento da popolazioni finite, quanto maggiore è l'utilizzazione di informazioni ausiliarie tanto minore diventa l'utilizzazione del «caso» nella scelta delle unità da campionare. Questo tema ha suscitato un animato dibattito negli anni '30; anni in cui si andavano definendo i fondamenti dell'inferenza statistica su popolazioni finite.

2. UTILIZZAZIONE DI INFORMAZIONI AUSILIARIE NELLA FASE DEL DISEGNO

Presso l'Istat, le informazioni disponibili vengono attualmente utilizzate a livello intermedio rispetto alle due situazioni estreme ap-

pena menzionate. Le indagini campionarie Istat vengono effettuate infatti utilizzando disegni di campionamento complessi che prevedono sia procedimenti di stadificazione che di stratificazione. In particolare, per quanto riguarda le indagini basate sulle famiglie, il disegno di campionamento è a due stadi con stratificazione delle unità di primo stadio (UP).

Le UP sono costituite dai comuni mentre quelle di secondo stadio (US) sono costituite dalle famiglie.

Le informazioni ausiliarie utilizzate in questa fase variano da indagine ad indagine. Quella comune a tutte le indagini basate sulle famiglie è, comunque, la dimensione demografica delle UP. Per l'indagine campionaria avente come scopo la rilevazione delle forze di lavoro vengono invece utilizzate, come informazioni ausiliarie per la stratificazione, le seguenti variabili:

1) Ampiezza demografica (calcolata utilizzando la popolazione residente);

2) Zona altimetrica delle PSU;

3) Attività economica prevalente;

4) Settori statistici.

Informazioni ulteriori sul disegno di campionamento sono reperibili in «Metodi e Norme» Istat, 1978.

Queste stesse variabili di stratificazione venivano utilizzate anche per altre indagini campionarie; esse sono state successivamente abbandonate poiché è stato sperimentato (Napolitano, Russo, Zannella, 1983; Russo, 1984) che, il guadagno in efficienza degli stimatori non variava significativamente utilizzando opportunamente soltanto la dimensione demografica delle UP invece di tutte e tre le variabili [1]. Un'ulteriore sperimentazione che conferma i precedenti risultati è stata effettuata da Zannella (1987) nell'ambito dei lavori della «Commissione Campioni».

Altre informazioni vengono inoltre utilizzate al fine di definire i domini territoriali di riferimento delle stime. Per quanto riguarda l'indagine sulle forze di lavoro, queste riguardano l'appartenenza delle unità della popolazione alle diverse regioni o province italiane.

Come è facile osservare, le variabili di tipo territoriale non si modificano sostanzialmente nel tempo, mentre le altre subiscono delle

variazioni temporali per cui necessitano di continui aggiornamenti per poter essere utilizzate al momento della predisposizione del disegno di campionamento.

Un ulteriore motivo per l'uso della dimensione demografica delle UP, sia nella fase del disegno che in quella della costruzione delle procedure di stima, è dato dal fatto che quest'ultima risulta sistematicamente aggiornata.

Un altro contributo di studio e sperimentazione nell'uso di informazioni disponibili nella fase del disegno è contenuto nel lavoro di Di Traglia e Russo (1984) e riguarda l'adozione di un disegno ppsx (disegno di campionamento e probabilità proporzionale ad una variabile X che, nel nostro caso, rappresenta l'ampiezza demografica delle UP).

L'insieme delle modalità assunte dalla variabile X sulle N unità della popolazione, rappresenta l'informazione disponibile utilizzata nella fase di costruzione delle probabilità di selezione delle stesse unità.

La metodologia studiata, basata in larga parte sulla proposta di Hartley, Rao e Cochran (1962), è stata poi applicata all'«Indagine sulle strutture ed i comportamenti familiari» (Istat, 1985). In tale occasione è stata anche predisposta una procedura informatica in linguaggio «speakeasy», per il calcolo delle probabilità di selezione PPSX e la successiva selezione delle unità di primo stadio. Ulteriori approfondimenti sono contenuti nel volume che riporta i risultati dell'indagine (Russo e Di Traglia, 1985). Per quanto riguarda invece le procedure di stratificazione un risultato relativo alla determinazione dei limiti ottimali degli strati è stato ottenuto da Falorsi (1984), utilizzando la tecnica dell'analisi in componenti principali e partendo da alcuni risultati ottenuti da Dalenius (1959).

3. UTILIZZAZIONE DI INFORMAZIONI AUSILIARIE NELLA FASE DI STIMA CON PARTICOLARE RIFERIMENTO ALLE INDAGINI PERIODICHE

Come è noto, l'uso opportuno di informazioni ausiliarie nella fase di costruzione delle procedure di stima consente di aumentare l'efficienza degli stimatori. Presso l'Istat, la metodologia di stima generalmente adottata è quella dello stimatore del rapporto post-

stratificato. Le informazioni ausiliarie sono rappresentate, per quanto riguarda le indagini basate sulle famiglie, dalla popolazione residente classificata secondo il sesso e le classi di età. Tali aspetti verranno trattati ampiamente nel Capitolo I del presente volume. È comunque da aggiungere che le informazioni ausiliarie nella fase di stima delle quantità di interesse della popolazione, devono essere sempre riferite al periodo in cui vengono effettuate le indagini. In tal senso tutte le informazioni utilizzate nelle indagini Istat possono essere definite «informazioni trasversali».

Esistono però alcune indagini effettuate periodicamente che forniscono informazioni sulla dinamica temporale di alcuni importanti fenomeni.

L'insieme dei risultati di tali indagini possono definirsi «informazioni longitudinali». Sono di questo tipo, ad esempio, le indagini sui consumi e sulle forze di lavoro.

Dalla letteratura internazionale, possono essere reperiti numerosi esempi di utilizzazione dell'insieme delle informazioni longitudinali disponibili, in fase di costruzione delle procedure di stima (Bureau of Census, 1978; Istituto Nacional de Estadistica, 1975).

A tal fine si ricorre all'impiego dello stimatore «composto», che può essere scritto, nella forma generale, nel seguente modo:

$$\hat{Y}_t = (1 - k)\,\hat{Y}'_t + k\left[\hat{Y}_{t-\Delta t} + (\hat{Y}'_t - \hat{Y}''_{t-\Delta t})\right] \qquad (1)$$
$$(0 \leq k \leq 1)$$

dove:

$\hat{Y}_t$ = stima del parametro di interesse nella popolazione al tempo t;

$\hat{Y}'_t$ = stima del parametro di interesse nella popolazione al tempo t ottenuta utilizzando l'intero campione;

$\hat{Y}_{t-\Delta t}$ = stima del parametro di interesse nella popolazione al tempo precedente $t - \Delta t$;

$\hat{Y}''_t$ = stima del parametro di interesse ottenuta utilizzando il sottocampione delle unità presenti sia al tempo t che al tempo $t - \Delta t$.

Come è facile notare la (1) è una combinazione lineare convessa tra la stima della quantità di interesse ottenuta con il campione

al tempo t ed una «previsione» ottenuta aggiornando la stima al tempo (t—Δt) con una stima del cambiamento avvenuto nel fenomeno $\hat{Y}_t$ nell'intervallo Δt.

Tale stimatore, al momento, non viene utilizzato in nessuna indagine Istat.

Tuttavia, presso il nostro Reparto, nel corso degli ultimi anni, sono stati già affrontati alcuni aspetti dei complessi problemi, metodologici ed operativi, connessi all'introduzione di tale stimatore.

4. ATTUALI FILONI DI RICERCA: DESCRIZIONE DEGLI ASPETTI METODOLOGICI ESSENZIALI

Le difficoltà nell'uso di questi stimatori derivano principalmente dal fatto che, per stimare la quantità $(\hat{Y}_t - \hat{Y}_{t-\Delta t})$ devono essere utilizzate le stesse unità campione sia al tempo t che al tempo t — Δt.

Negli ultimi anni, però, sia per l'introduzione nell'ambito dell'inferenza da popolazioni finite dei modelli statistico-matematici, sia per i risultati incoraggianti ottenuti, nell'uso delle informazioni longitudinali nella fase di stima, da istituti di statistica di altri Paesi, è stato affrontato anche presso l'Istat il problema di come queste informazioni possano essere introdotte nelle nostre procedure di stima. Un approfondimento teorico relativo al problema dell'uso di informazioni derivanti da precedenti indagini è reperibile in Di Traglia e Falorsi (1987). In questo lavoro viene ripresa una proposta di Scott e Smith (1978) — i quali seguendo l'ottica bayesiana sviluppano una metodologia per la costruzione di uno stimatore composto — e trasferita nell'ambito dell'approccio classico utilizzando modelli di superpopolazione.

I risultati che emergono riguardano essenzialmente il fatto che il cambiamento delle unità campionate nei diversi istanti di tempo, non influisce sulla stima delle variazioni dei parametri della popolazione, nei diversi intervalli temporali. Questo perché la variabilità del carattere oggetto di indagine, dovuta al cambiamento delle unità campione nei diversi tempi di campionamento, viene inglobata nel modello di superpopolazione insieme alla variabilità temporale residua del fenomeno oggetto di indagine.

Tale variabilità viene poi utilizzata nel calcolo del peso k della combinazione lineare convessa tra stima campionaria e «previsione» temporale. La formula generale dello stimatore è data da:

$$\hat{Y}_t = (1 - k)\,\hat{Y}'_t + k\,\hat{u}_t \qquad (2)$$

dove:

$$k = \frac{\hat{S}^2_t}{V\left[\hat{Y}'_t / \hat{Y}'_{t-\Delta t}\right]} \qquad (3)$$

$\hat{S}^2_t$ è una stima della varianza campionaria della stima ottenuta dal campione al tempo t.

Poiché $\hat{Y}'_t$ è una stima campionaria di Y'_t condizionalmente a $\hat{Y}_{t-\Delta t}$ si ha:

$$V\left[\hat{Y}_t/\hat{Y}_{t-\Delta t}\right] = V\left[\hat{Y}'_t/\hat{Y}_{t-\Delta t}\right] + \hat{S}^2_t \qquad (4)$$

La formula (2) presuppone una dinamica temporale del «parametro di popolazione» (oggetto di stima) di tipo autoregressivo del primo ordine (processo markoviano). Nella realtà tale ipotesi può risultare restrittiva.

È per questo motivo che, se pur nell'ambito della soluzione di un problema diverso (stime preliminari), viene presentato un modello di stima che tiene conto di tutte le componenti (trend, ciclo e stagionalità) solitamente presentati nella dinamica dei fenomeni (economici e socio-demografici) generalmente oggetto di studio nelle indagini Istat (Russo e Di Traglia, 1988).

Nel lavoro citato viene studiata, al fine di aumentare la precisione degli stimatori (quando non è ancora disponibile l'intero campione) una strategia mista.

L'indagine presa in considerazione, per riferire l'applicazione dei risultati dello studio, è stata quella delle forze di lavoro, e lo scenario ipotizzato quello in cui, ad un certo istante di tempo t (successivo alla spedizione dei questionari ai comuni campione), sono disponibili presso l'Istat soltanto i questionari relativi ad m degli n comuni campione con $m < n$. Il problema da risolvere è quello di ricavare delle stime sufficientemente stabili per le principali variabili ogget-

to di studio. Si ritiene utile fornire, in questa sede, alcune indicazioni metodologiche non chiaramente espresse, per motivo di spazio, nel lavoro citato.

La proposta ivi contenuta si articola in due fasi. Nella prima l'insieme (S) delle n unità campione viene partizionato in otto sottoinsiemi ottenuti attraverso la seguente metodologia.

Indichiamo con K il numero complessivo delle variabili prese in considerazione nell'indagine, ed applichiamo, alle n UP campione, un algoritmo di classificazione ottenendo H gruppi sulla base della loro vicinanza nello spazio K-dimensionale.

Per ciascun gruppo è possibile, fissando una certa soglia, definire un sottogruppo di unità campione in grado di «rappresentare» al meglio le altre (all'interno di ciascun gruppo). Queste vengono chiamate unità rappresentative ed indicate con S^h_r. Le rimanenti sono quelle non rappresentative ed indicate con S^h; si avrà:

$$\bigcup_{h=1}^{H} (S^h \cup S^h_r) \cup S^h_a = S \qquad (5)$$

dove S^h_a è l'insieme delle UP autorappresentative. Fissato un instante di tempo t, le unità disponibili vengono indicate con d, quelle non disponibili con $\bar{d}$.

La seguente tabella può essere utile a chiarire meglio il problema:

	S_r	S_a	$S(r)$	$S(\bar{r})$
d	S_{dr}	S_{da}	$S_d(r)$	$S_d(\bar{r})$
$\bar{d}$	$S_{\bar{d}r}$	$S_{\bar{d}a}$	$S_{\bar{d}}(r)$	$S_{\bar{d}}(\bar{r})$

(6)

dove, oltre ai simboli già noti, compaiono $S_d(r)$, $S_{\bar{d}}(r)$, $S_d(\bar{r})$ $S_{\bar{d}}(\bar{r})$ che derivano dalla partizione delle unità non rappresentative sulla base delle disponibilità (al tempo t) loro e di quelle che le rappresentano.

Per quella parte della popolazione le cui unità campione cadono nell'insieme

$$Q = \left\{S_{dr} \cup S_{da} \cup S_{d(r)} \cup S_d(\bar{r}) \cup S_{\bar{d}}(r)\right\} \qquad (7)$$

viene utilizzato lo stimatore del rapporto post-stratificato corrente-

mente adottato, mentre per la parte di popolazione le cui unità campione appartengono all'insieme:

$$\{M\} = \{S_{\bar{d}r} \cup S_{\bar{d}a} \cup S_{\bar{d}}(\bar{r})\} \tag{8}$$

viene utilizzato un «predittore» del tipo:

$$\hat{T}_{Mi} = \sum_{m \in M} \hat{T}_{mi}(t-1) + (\bar{\Phi}_m B + \beta_m B^s - \bar{\Phi}_m \beta_m B^s) \Delta \hat{T}_{mi} \tag{9}$$

dove $\hat{T}_{Mi}$ rappresenta una stima del totale della i-esima variabile oggetto d'indagine, relativamente all'insieme degli individui della popolazione che appartengono ad $\{M\}$.

Viene, infine, nello stesso lavoro affrontato il problema della distribuzione dei tempi di ritorno presso l'Istat (dei questionari inviati alle UP per la compilazione) al fine della scelta del momento in cui calcolare le stime preliminari.

Un ulteriore approccio al problema dell'utilizzazione delle informazioni ausiliarie scaturisce dalla considerazione dell'uso congiunto di informazioni «trasversali» e «longitudinali». Come è venuto meglio evidenziandosi, sembra che l'approccio basato sui modelli di superpopolazione consenta di sfruttare maggiormente le informazioni disponibili.

In particolare sembra più fruttuoso tale approccio se queste informazioni si riferiscono a risultati (sulle stesse variabili) riferite alle indagini precedenti.

L'utilizzazione delle informazioni longitudinali (attraverso modelli di superpopolazione) non esclude comunque la possibilità di utilizzare anche informazioni trasversali seguendo la metodologia classica. In tale ambito si ottengono stimatori abbastanza semplici da poter essere applicati nelle indagini su larga scala quali quelle Istat, (Di Traglia, 1988).

La formula, in tale occasione proposta, sfrutta l'ipotesi che il parametro di popolazione (oggetto di stima) abbia una dinamica temporale descrivibile da un modello autoregressivo con una componente di trend, una stagionale ed una erratica. Utilizzando i noti operatori Δ e B, la componente erratica, per ciascuna stima che compare nelle tavole (matrici) delle stime campionarie $\hat{N}_{ij}(t)$ al tempo t risulta:

$$u_{ij}(t) = (1 - \bar{\Phi}_{ij} B)(1 - \beta_{ij} B^{s_{ij}}) \Delta \hat{N}_{ij}(t)$$

da cui: (10)

$$\tilde{N}_{ij}(t-1) = \hat{N}_{ij}(t) + (\bar{\Phi}_{ij} B + \bar{\Phi}_{ij} B^{s_{ij}} - \bar{\Phi}_{ij} \beta_{ij} B)^{(s_{ij}+1)} \Delta \hat{N}_{ij}(t)$$

Supponendo che il periodo di riferimento dell'indagine campionaria sia T + 1 e che inoltre siano conosciuti a livello di popolazione le quantitià $N_{.j}(T+1)$ $j = 1,2,.....H$ ed indicando con $n_{ij}(T+1)$ i dati disponibili, è possibile utilizzare lo stimatore del rapporto sia per la previsione tramite modello che per la stima campionaria. Si ha infatti:

$$\tilde{N}^{R}_{ij} = \frac{\tilde{N}_{ij}}{\sum_i \tilde{N}_{ij}} N_{.j}$$

$$\hat{N}_{ij} = \frac{n_{ij}}{n_{.j}} N_{.j} \qquad (11)$$

Combinando le due stime sotto il vincolo $\beta(N) = 0$ si ottiene:

$$\hat{N}^{*}_{ij}(T+1) = \left[1 - \beta(n)\right] \hat{N}_{ij}(T+1) + \beta(n) \tilde{N}^{R}_{ij}(T+1) \qquad (12)$$

dove:

$$\beta(n) = \frac{N - n}{(\bar{\Phi}_{ij} + \beta_{ij} - \bar{\Phi}_{ij} \beta_{ij}) N} \qquad (13)$$

ed $\hat{N}^{*}_{ij}$ rappresenta la stima del totale del carattere ottenuta combinando il modello (10) con gli stimatori (11).

È ancora possibile «migliorare» lo stimatore (12) utilizzando i legami interni delle serie storiche attraverso l'analisi delle correlazioni incrociate.

Bisogna dire comunque che, nel campo dell'utilizzazione delle informazioni riguardanti la dinamica temporale dei fenomeni oggetto di indagine, i contributi Istat sono, a tutt'oggi, prevalentemente di natura teorica. Non è stato infatti ancora possibile effettuare alcuna sperimentazione. Si ritiene però, che data l'attualità del problema ed i possibili benefici in termine di efficienza degli stimatori (e quindi delle possibilità di diminuire — a parità di efficienza — la numerosità campionaria), dovranno essere effettuati degli esperimenti possibilmente su dati derivanti da indagini reali.

5. CONCLUSIONI E FUTURI ITINERARI DI RICERCA

L'approccio al problema della stima da popolazioni finite, basato sull'uso dei modelli di superpopolazione, come già accennato, consente un'utilizzazione maggiore dell'informazione disponibile. Questo fatto comunque non avviene solo in presenza di informazioni derivanti da indagini correnti e ripetute periodicamente ma anche quando le informazioni derivano da indagini trasversali. A tal fine, presso il Progetto «Studio dei campioni» è stato effettuato uno studio teorico su una metodologia di stima che, nel caso di un disegno stratificato, sfrutta anche l'informazione contenuta nella relazione tra strati diversi (Russo e Di Traglia, 1988). Il risultato, nel caso di stimatori lineari, è un modello lineare dinamico per ciascun parametro oggetto di stima. In questa direzione, inoltre, sono ancora da sperimentare altri risultati riguardanti i modelli, comunemente usati nell'analisi dei dati (log-lineari, logistici, ecc), nelle metodologie di stima per popolazioni finite. Un ulteriore campo di approfondimento sperimentale è l'interazione tra disegno e modello per le scelte di «strategie di campionamento» ottimali.

Relativamente a quest'ultimo punto sono da approfondire gli aspetti legati all'uso del coefficiente di correlazoine intraclasse e del deft nell'ambito dell'approccio basato sui modelli statistico-matematici. Ulteriori sperimentazioni verranno effettuate sull'uso dei modelli di presentazione della varianza delle stime, sia per le scelte di strategie di campionamento ottimali, sia al fine di ottenere una stima più precisa della varianza stessa.

NOTE

(1) Le dimensioni degli strati vengono costruite in modo da rendere il più omogenee possibile, le unità di primo stadio universo, negli stessi strati, rispetto alla variabile oggetto d'indagine. Per quanto riguarda l'indagine Forze di lavoro, questa viene, invece, utilizzata nel suddividere le UP in due strati; comuni inferiori a 20.000 abitanti e comuni superiori a 20.000 abitanti.

RIFERIMENTI BIBLIOGRAFICI

COCCIA G., RUSSO A., FALORSI P., D'ANGIOLINI G. (1987) «*Una metodologia per la valutazione degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo del disegno di campionamento nell'indagine sulle forze di lavoro*». Seminario su «Forze di lavoro: Disegno dell'indagine ed analisi strutturale», Bressanone.

FALORSI P.D. (1984) «*Sulla stratificazione delle unità di primo stadio nelle indagini campionarie Istat presso le famiglie*» Atti della XXXII Riunione Scientifica della Società Italiana di Statistica, Sorrento.

Istat (1978), «*Rilevazione campionaria delle forze di lavoro*», Metodi e Norme N. 15, Roma.

Istat (1988), «*Indagine sugli sport e sulle vacanze*», Metodi e Norme N. 2, Roma.

NAPOLITANO P., RUSSO A., ZANNELLA F. (1980) «*Calcolo, presentazione ed analisi degli errori di campionamento dell'indagine Istat sulle condizioni di salute della popolazione e sul ricorso ai servizi sanitari*». Atti del Convegno SIS, Trieste.

DI TRAGLIA M. (1987) «*Teoria della superpopolazione e campionamento da popolazioni finite*», Tesi del dottorato di ricerca in statistica Metodologica, Roma.

DI TRAGLIA M. (1988) «*Post-stratificazione e modelli di superpopolazione*». Atti della XXXIV Riunione SIS, Siena.

DI TRAGLIA M. (1988) «*The use of longitudinal data to obtain estimates in the sampling from finite populations*», IAOS, Roma.

DI TRAGLIA M., FALORSI S. (1987) «*Approccio model-based nel campionamento in tempi successivi per piccole aree*». Atti del Convegno SIS su informazione ed analisi statistica per aree regionali e sub-regionali, Perugia.

DI TRAGLIA M., RUSSO A. (1988) «*Model-based approach to estimates from finite populations using multiway data*». Presentato al Convegno «Multiway '88», Roma.

DI TRAGLIA M., RUSSO A. (1989) «*A model-based approach to estimates from finite populations*». In corso di pubblicazione in Proceedings of 47th Session of the ISI, Parigi.

RUSSO A., DI TRAGLIA M. (1984) «*Metodologia per il riporto dei dati all'universo e per il calcolo degli errori di campionamento nell'indagine Istat sulle strutture ed i comportamenti familiari*». Atti della XXXII Riunione Scientifica della Società Italiana di Statistica, Sorrento.

RUSSO A. (1984) «*Calcolo ed analisi degli errori di campionamento nell'indagine Istat sulle vacanze e gli sport degli italiani - 1983*». Atti della XXXII Riunione Scientifica della Società Italiana di Statistica, Sorrento.

RUSSO A., DI TRAGLIA M. (1985) «*Disegno di campionamento*», Indagine sulle strutture e comportamenti familiari», Istat, Roma.

RUSSO A., DI TRAGLIA M. (1988) «*A methodology to obtain preliminary estimates from repeated surveys*», Conference of European Statisticians - Seminar on Statistical methodology, ONU, Ginevra.

RUSSO A., FALORSI P., GIOVANI P. (1989) «*Valutazione dell'efficienza delle stime ed analisi critica del disegno di campionamento adottato nell'indagine sulle forze del lavoro*». Apparirà sul Volume «Convegno FOLA '89».

SCOTT A., SMITH T.M.F. (1974) «*Analysis of repeteated su using time series methdos*». JASA, 59, 674-678.

TOM S.M. (1987) «*Analysis of repeated surveys using dinamic linear model*». ISI, 55, 63-73.

ZANNELLA F. (1989) «*Metodologia, programmi ed esperimenti relativi alla progettazione di una procedura generale per la stratificazione dei comuni*». Istat, «Commissioni di studio sui campioni».

CALCOLO E PRESENTAZIONE DEGLI ERRORI DI CAMPIONAMENTO

di *Giuliana Coccia*

1. INTRODUZIONE

La letteratura statistica distingue gli errori che possono presentarsi in ogni indagine campionaria in due parti:

— errori di campionamento, dovuti alla natura parziale della rilevazione;

— errori di misura (o di risposta), derivanti da numerosi e spesso incontrollabili fattori di disturbo (es. elenchi base errati, domande mal formulate, mancate risposte, ecc.).

L'esistenza di tali errori induce lo statistico, al fine di minimizzarne la portata, a prendere decisioni e provvedimenti che investono da una parte gli aspetti del disegno di campionamento (stratificazione, particolari procedure di stima) e dall'altra le principali operazioni sia preparatorie che esecutive riguardanti l'indagine (redazione del modello, istruzione ai rilevatori, ecc.).

D'altra parte l'esperienza spesso mostra che, completata la rilevazione ed effettuate le operazioni tradizionali di revisione ed analisi critica dei dati, i risultati dell'indagine vengono considerati pronti per essere passati agli utilizzatori, assumendo che gli errori campionari e gli errori di misura siano di entità limitata e comunque tali da non compromettere seriamente la qualità dei risultati.

Tale assunzione dovrebbe essere sempre verificata valutando l'affidabilità dei risultati, in funzione del livello di precisione e di accuratezza riguardante tutti gli errori di misura.

È auspicabile, quindi, che le relazioni finali sui risultati delle indagini contengano un capitolo speciale in cui siano descritti il disegno di campionamento adottato, le procedure di stima, gli errori campionari e di misura ed alcune statistiche ausiliarie utili nell'inferenza e nell'analisi statistica.

In questa ottica l'Istat, negli ultimi 10 anni, ha portato avanti un programma di studi per affrontare sistematicamente e scientificamente il calcolo degli errori di campionamento.

Nell'ambito di questo programma sono stati affrontati anche i problemi concernenti la presentazione «sintetica» degli errori campionari. Negli ultimi anni, a tale argomento è stata data viva attenzione soprattutto dai grossi centri di informazione statistica a livello internazionale, che, come l'Istat, hanno il compito di effettuare indagini su larga scala finalizzate alla produzione di un numero molto elevato di stime. L'utilizzazione di metodi per la presentazione sintetica è di notevole utilità in quanto consente di ridurre sensibilmente i tempi ed i costi connessi al calcolo degli errori campionari.

2. CALCOLO DEGLI ERRORI DI CAMPIONAMENTO

All'inizio degli anni '80 il problema del calcolo degli errori campionari veniva affrontato in un'ottica basata sull'utilizzazione di espressioni «esatte» delle varianze campionarie. Ciò comportava la messa a punto di programmi informatici specifici per ciascuna indagine (Napolitano e Russo, 1979; Napolitano e Russo, 1980; Russo, 1982; Di Traglia e Russo, 1982).

Per superare le difficoltà derivanti dalla predisposizione di metodologie e programmi specifici per ciascuna indagine, si è cominciato a sperimentare il programma generalizzato CLUSTERS, predisposto dallo Staff tecnico della World Fertility Survey (Verma e Pearce, 1978).

Attualmente l'Istat utilizza tale programma per il calcolo delle varianze di campionamento di tutte le indagini condotte sulle famiglie.

Il CLUSTERS, basato su un procedimento approssimato, usa il metodo dello sviluppo in serie di Taylor di una funzione e si basa sulle seguenti assunzioni:

a) che da ogni strato siano estratte due o più unità primarie;
b) che tali estrazioni siano indipendenti.

Tale programma consente il calcolo dell'errore di campionamento — assoluto e relativo — e di altre statistiche quali il coefficiente di correlazione intraclasse, l'effetto del disegno di campionamento, l'intervallo di confidenza ed il numero medio di unità rilevate per unità primaria.

Per illustrare la metodologia alla base del programma in questione, faremo riferimento ad una stima rapporto in quanto questa

comprende come casi particolari la stima di frequenze assolute, di percentuali e di medie.

Indichiamo con Y_{his} il valore del carattere y osservato sull'unità finale s rilevata nell'unità primaria i dello strato h e con K_{his} il coefficiente di ponderazione attribuito a tale unità. La stima del totale del carattere y relativo alla generica unità primaria i è fornito da:

$$\hat{Y}_{hi} = \sum_{s \epsilon m_{hi}} K_{his} Y_{his} \tag{1}$$

in cui m_{hi} è l'insieme delle unità finali campionate nell'unità primaria i dello strato h.

Conseguentemente la stima per lo strato h è data da:

$$\hat{Y}_h = \sum_{i \epsilon n_h} \hat{Y}_{hi} \tag{2}$$

in cui n_h è l'insieme delle unità primarie campione nello strato h.

Il totale generale si ottiene come somma degli strati h:

$$\hat{Y} = \sum_{h \epsilon H} \hat{Y}_h \tag{3}$$

Espressioni analoghe possono essere scritte per un altro carattere x; pertanto la stima del rapporto fra due variabili è data da:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \tag{4}$$

Sotto l'ipotesi che le n_h unità primarie siano estratte in ciascuno strato in modo indipendente, la stima della varianza di $\hat{R}$ è espressa da:

$$\hat{V}(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h \epsilon H} \frac{n_h}{n_h - 1} \sum_{i \epsilon n_h} \left(\hat{D}_{hi} - \frac{1}{n_h} \sum_{i \epsilon n_h} \hat{D}_{hi}\right)^2 \tag{5}$$

in cui si è posto:

$$\hat{D}_{hi} = \hat{Y}_{hi} - \hat{R}\,\hat{X}_{hi} \tag{6}$$

Posto inoltre:

$$\hat{\bar{Y}} = \hat{R}\,X \tag{7}$$

in cui X è il totale del carattere x, la stima della varianza della stima del rapporto $\hat{\hat{Y}}$ è data da:

$$\hat{V}(\hat{\hat{Y}}) = \sum_{h \in H} \frac{n_h}{n_h - 1} \sum_{i \in n_h} (\hat{D}_{hi} - \frac{1}{n_h} \sum_{i \in n_h} \hat{D}_{hi})^2 \qquad (8)$$

Ciò premesso, riteniamo utile descrivere l'applicazione di tale metodologia al caso delle indagini Istat basate sulla famiglie.

In tali indagini vengono adottati disegni di campionamento a due stadi con stratificazione delle unità primarie (comuni). Nell'ambito di ciascuna regione geografica i comuni sono ripartiti in due gruppi:

— dominio autorappresentativo (AR) costituito dai comuni di dimensione demografica più elevata; tali comuni vengono tutti inclusi nel campione e quindi costituiscono ciascuno uno strato a sé stante;

— dominio non autorappresentativo (NAR) formato dai rimanenti comuni, i quali vengono suddivisi in strati; da ogni strato, in genere, viene estratto un solo comune con probabilità proporzionale alla sua dimensione demografica.

Tenendo presente le caratteristiche del disegno di campionamento ed indicando con $\hat{\hat{Y}} = \hat{\hat{Y}}_{AR} + \hat{\hat{Y}}_{NAR}$ la stima del totale del carattere y riferito ad un dato dominio territoriale, la varianza di $\hat{\hat{Y}}$ risulta espressa dall'espressione seguente:

$$V(\hat{\hat{Y}}) = V(\hat{\hat{Y}}_{AR}) + V(\hat{\hat{Y}}_{NAR}) \qquad (9)$$

L'adattamento formale della (8) per il dominio autorappresentantivo porta immediatamente alla formula:

$$\hat{V}(\hat{\hat{Y}}_{AR}) = \sum_{h \in H_1} \frac{m_h}{m_h - 1} \sum_{i \in m_h} (\hat{D}_{hi} - \frac{1}{m_h} \sum_{i \in m_h} \hat{D}_{hi})^2 \qquad (10)$$

in cui m_h è l'insieme delle famiglie campione (unità primarie) nello strato h, H_1 è l'insieme degli strati definiti in AR ed inoltre:

$$\hat{D}_{hi} = \hat{Y}_{hi} - \hat{R}\,\hat{X}_{hi} \qquad (11)$$

Per il dominio NAR, avendo una sola unità primaria campione in ogni strato, non è possibile definire una stima consistente di $V(\hat{\hat{Y}}_{NAR})$.

In tali situazioni il programma in questione prevede la possibilità o di accoppiare automaticamente due UP consecutive, o di definire nuovi strati raggruppando, mediante una scelta ragionata, due o più UP appartenenti a strati diversi; nel nostro caso viene adottata la prima alternativa. In questa circostanza l'espressione della varianza alla base del CLUSTERS assume la forma:

$$\hat{V}(\mathring{\hat{Y}}_{NAR}) = \sum_{h \in H_2/2} \frac{n_h}{n_h - 1} \sum_{i \in n_h} (\hat{D}_{hi} - \frac{1}{n_h} \sum_{i \in n_h} \hat{D}_{hi})^2 \qquad (12)$$

in cui $n_h = 2$ sono i comuni campione (unità primarie) nel «superstrato» costruito, H_2 è l'insieme degli strati definiti in NAR, D_{hi} è stato definito nell'espressione (6).

Peraltro, gli studi teorici (Russo, 1985; Coccia, Falorsi, Russo, 1986) finalizzati alla determinazione dell'espressione della varianza in contesti campionari che prevedono l'estrazione di una sola UP per strato, hanno condotto ad espressioni che hanno la stessa struttura formale della (12). Tali studi sono stati sviluppati in un'ottica fondata sulla generalizzazione di un criterio «collapsed strata» suggerito da Cochran (1977), per un disegno campionario molto più semplice di quelli alla base delle indagini Istat.

Nel quadro di questi studi riteniamo opportuno segnalare un lavoro (Falorsi 1988) che affronta il problema della distorsione di $\hat{V}(\mathring{\hat{Y}}_{NAR})$ nelle due componenti: la prima relativa all'accoppiamento degli strati; la seconda dovuta al fatto che $\mathring{\hat{Y}}_{NAR}$ è uno stimatore distorto.

Si fà tuttavia presente che gli errori di campionamento calcolati si riferiscono al dominio territoriale regionale, costituito dai due domini AR e NAR. In tali circostanze, la componente distorsiva dovuta all'utilizzazione degli strati collapsed può ritenersi trascurabile; per quanto concerne la seconda componente, le esperienze condotte in altri Paesi mostrano che essa è di lieve entità.

3. PRESENTAZIONE DEGLI ERRORI DI CAMPIONAMENTO

Un'informazione completa sul livello di precisione dei risultati richiederebbe la specificazione degli errori di campionamento di tutte le stime pubblicate. Tuttavia ciò non è proponibile in quanto comporterebbe un elevato numero di elaborazioni informatiche, un'ap-

pesantimento del volume riportante i dati dell'indagine e di conseguenza tempi e costi molto elevati. Tali difficoltà hanno offerto lo spunto per introdurre idonei modelli che consentono di esporre in forma concisa gli errori campionari.

I modelli proposti e ritenuti validi a tutt'oggi si distinguono secondo due diversi approcci: uno basato sui modelli regressivi (Bean, 1970), l'altro sull'utilizzazione dell'effetto complessivo del disegno di campionamento o deft (Verma, 1982).

L'Istat ha dedicato un'attenzione crescente al metodo basato sui modelli regressivi, sviluppando un filone di ricerca sia di carattere metodologico (Russo, 1985; Falorsi e Russo, 1985, Russo, 1987) sia di carattere empirico (Di Traglia e Russo, 1982; Zannella, 1982; Di Traglia e Russo, 1985; Coccia, Falorsi e Russo, 1986).

Ricordiamo che il metodo dei modelli regressivi si fonda sulla relazione esistente fra stime ed errori percentuali di campionamento, in base alla quale l'errore relativo decresce all'aumentare della stima. Le funzioni che si sono dimostrate più idonee a descrivere la suddetta relazione hanno la forma:

$$\frac{\hat{V}(\hat{Y})}{\hat{Y}^2} = a + \frac{b}{\hat{Y}} \tag{13}$$

$$\log \frac{\hat{V}(\hat{Y})}{\hat{Y}} = a + b \log \hat{Y} \tag{14}$$

Esaminiamo ora l'utilizzazione dei modelli sopra descritti nelle indagini Istat basate sulle famiglie. A tal fine supponiamo che una data indagine abbia fornito s stime, indicate con $G_s = (\hat{Y}_1, ..., \hat{Y}_i, ... \hat{Y}_s)$; l'approccio che viene adottato può essere sintetizzato dai seguenti steps:

a) suddivisione delle stime oggetto di indagine in t sottogruppi, sotto il vincolo che le stime appartenenti a ciascun sottogruppo abbiano un deft approssimativamente costante. In pratica la suddivisione in esame tiene conto dell'esperienze passate e di alcune indicazioni utili allo scopo, quali riferire le stime allo stesso livello territoriale ed alle stesse caratteristiche demografiche ed economiche;

b) estrazione da ciascun sottogruppo di un determinato numero di stime; questo è uno dei momenti più importanti in quanto le stime prescelte devono coprire un ampio «range»;

c) calcolo delle varianze relative di tutte le stime prescelte mediante il programma CLUSTERS;

d) stima dei parametri incogniti a e b, che caratterizzano il modello di regressione, con il metodo dei minimi quadrati;

e) determinazione dei valori degli errori campionari, mediante le funzioni stimate, per particolari livelli di stime.

Nei volumi riportanti i risultati delle indagini, ai fini della presentazione degli errori interpolati, l'ottica attuale è quella di pubblicare, per ciascuna regione geografica, una tabella in cui vengono riportati diversi livelli di stima ed i corrispondenti errori percentuali distintamente per le due sottoclassi maschi e femmine. Inoltre riteniamo opportuno segnalare che viene altresì pubblicata una tabella riportante i valori dei coefficienti a e b, attraverso i quali gli utilizzatori possono determinare gli errori per valori di stime diversi da quelli riportati nella prima tabella.

Il secondo approccio basato sull'effetto del disegno di campionamento e di alcune sue componenti, è stato sviluppato negli ultimi anni (Verma, Scott e O'Muirchaertaigh, 1980; Verma, 1982).

A tale scopo ricordiamo che il deft è un fattore che sintetizza le varie complessità del disegno di campionamento, come ad esempio quelle relative alla stratificazione, all'introduzione dei pesi variabili, alla post-stratificazione, ecc. Di conseguenza la varianza di un campione complesso, $V_c(\hat{Y})$, può essere espressa come prodotto della varianza di un campione casuale semplice delle stesse dimensioni in termini di unità finali, $V_s(\hat{Y})$, per il fattore deft, che misura l'efficienza del disegno complesso rispetto al disegno casuale semplice. Questa relazione può essere definita:

$$\sqrt{V_c(\hat{Y})} = \sqrt{V_s(\hat{Y})}\,\text{deft} \qquad (15)$$

In questa nota, per ragioni di tempo, non illustriamo in modo dettagliato l'insieme dei metodi proposti nell'ambito di questo approccio; tuttavia descriviamo l'idea fondamentale che ne sta alla base.

Partendo dall'espressione (15) si procede nel modo seguente:

a) scelta di alcuni gruppi fondamentali di variabili oggetto di indagine;

b) calcolo del deft medio nell'ambito di ciascun gruppo;

c) determinazione degli errori di campionamento riferiti ad un campione casuale semplice per diversi livelli di stima e di numerosità campionaria;

d) determinazione dell'errore di campionamento complesso per una generica stima, moltiplicando il deft medio del gruppo, al quale si può attribuire la stima in oggetto, per l'errore campionario della stessa sotto ipotesi di campionamento casuale semplice.

Ulteriori approfondimenti su tale approccio possono trovarsi nei lavori di Cohen (1982); Russo e Zonno (1987).

Tale metodologia è stata sperimentata per la prima volta all'Istat nell'ambito della presentazione degli errori di campionamento dell'indagine sugli sports e vacanze (1985). In particolare per tale indagine sono state adottate entrambe le metodologie; si è quindi proceduto ad un confronto tra i due tipi di valori interpolati calcolando le differenze medie assolute relative fra le varianze osservate e quelle interpolate, in particolare sono stati utilizzati i seguenti indici:

$$\bar{A}_1 = \sum_{t \epsilon T} \frac{|V(\hat{Y}) - V_r(\hat{Y})|}{n\, V(\hat{Y})} \qquad \bar{A}_2 = \sum_{t \epsilon T} \frac{|V(\hat{Y}) - V_d(\hat{Y})|}{n\, V(\hat{Y})} \qquad (16)$$

in cui t è l'insieme delle stime oggetto di indagine selezionate per il calcolo degli errori campionari, $V(\hat{Y})$ sono le varianze osservate, $V_r(\hat{Y})$ e $V_d(\hat{Y})$ sono le varianze interpolate rispettivamente mediante il metodo dei modelli regressivi e mediante il modello basato sul deft.

I valori ottenuti con l'indagine $\bar{A}_1$ sono risultati, in quasi tutte le regioni, più elevati di quelli ottenuti con $\bar{A}_2$; conseguentemente il metodo che utilizza il deft sembra essere una strategia più attendibile per stimare le varianze, conformemente a quanto sperimentato in altri Paesi.

4. CONSIDERAZIONI FINALI

Le considerazioni svolte nei paragrafi precedenti sono servite ad illustrare gli aspetti concettuali, metodologici ed operativi nell'ambito del calcolo e della presentazione degli errori di campionamento; tuttavia, il discorso sull'intera problematica è complesso e va ulteriormente approfondita.

È stato evidenziato che l'Istat ha abbandonato l'utilizzazione di espressione «esatte» per il calcolo delle varianze campionarie ed attualmente utilizza il programma CLUSTERS, basato sul metodo approssimato dello sviluppo in serie di Taylor; tale metodo è stato ampiamente sperimentato e può ritenersi soddisfacente.

Negli altri Paesi in special modo negli USA vengono frequentemente adottati altri due metodi approssimati noti in letteratura sotto il nome di:

— metodo delle replicazioni ripetute bilanciate (BRR);
— metodo jack-knife.

Tuttavia tali metodi hanno lo svantaggio, che, per ottenere livelli di precisione adeguati necessitano di un numero consistente di replicazioni che implicano, tenuto conto della complessità delle indagini effettive, costi e tempi tecnici di elaborazioni informatiche molto elevate.

I continui progressi in campo informatico e gli studi metodologici, sempre più numerosi nel corso di questi ultimi anni, fanno ritenere legittima l'attesa che nel prossimo futuro vengano messi a punto procedimenti che consentano di ridurre il numero di replicazioni.

Gli studi teorici condotti al fine di comparare le proprietà statistiche di cui godono i tre metodi sopra citati, non consentono di privilegiare uno dei tre.

Al momento attuale la scelta viene effettuata, pertanto, solamente sulla base di considerazioni di natura pratica.

RIFERIMENTI BIBLIOGRAFICI

BEAN J.A. (1970), *Estimation and Sampling Variance in the Health Interview Survey*, Vital and Health Statistics, Serie 2, N. 38.

COHEN S.B. (1982), *Comparison of design effect and relative variance curve strategy for variance estimation from complex survey data*, Annual Meeting of the American Pubblic Health Association.

COCCIA G., FALORSI P.D., RUSSO A. (1986), *Indagine speciale sulla lettura e su altri aspetti del tempo libero, 1984: Disegno di campionamento, calcolo e presentazione degli errori campionari*. Note e relazioni, 1986, n. 3, Istat, Roma.

COCCIA G., D'ANGIOLINI G., FALORSI P.D. e RUSSO A., (1987), *Una metodologia per la valutazione degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo del disegno di campionamento dell'indagine sulle forze di lavoro*, Atti del Seminario su: «Forze di lavoro: disegno dell'indagine ed analisi strutturali», Dipartimento di Scienze Statistiche - Università di Padova, Bressanone.

COCCIA G., FALORSI P.D., RUSSO A. (1988), *Indagine sugli sport e le vacanze degli italiani nel 1985*. Disegno di campionamento, procedura di stima, calcolo e presentazione degli errori campionari, Note e relazioni n. 2, Istat, Roma.

COCCIA G., FALORSI P.D., RUSSO A. e BOTTA M. (1988), *Indagine sugli sport e vacanze: gli sport degli italiani del 1985*. Disegno di campionamento, procedura di stima, calcolo e presentazione degli errori campionari, Note e relazioni, 1988, n. 3 Istat, Roma.

COCHRAN W.G. (1977), *Sampling Tecniques*, Wiley, New York.

FALORSI P.D. (1988), *Un metodo per la stima della varianza campionaria nei campioni a due stadi con una unità primaria per strato*, Atti della XXXIV Riunione Scientifica, SIS, Siena.

NAPOLITANO P., RUSSO A. e ZANNELLA F. (1983), *Calcolo, presentazione ed analisi degli errori di campionamento dell'indagine Istat sulle condizioni di salute della popolazione italiana e sul ricorso ai servizi sanitari*, Atti del Convegno della SIS, Trieste.

RUSSO A. e DI TRAGLIA M. (1982), *Distribuzione per età della popolazione scolastica, anno 1978-79: grado di attendibilità dei risultati*, Supplemento al Bollettino di Statistica, n. 25, Istat, Roma.

RUSSO A. (1984), *Calcolo ed analisi degli errori di campionamento nelle indagini Istat sulle vacanze e gli sport degli italiani, anno 1983*, Atti del XXXII Riunione della SIS, Sorrento.

RUSSO A. (1984), *Indagine sulle vacanze, i viaggi e gli sport degli italiani nel 1982: piano della rilevazione campionaria degli errori di campionamento*, Supplemento al Bollettino Mensile di Statistica n. 15, Istat, Roma.

RUSSO A. e DI TRAGLIA M. (1985), *Indagine sulle strutture ed i comportamenti familiari: Disegno di campionamento calcolo e presentazione degli errori campionari*, Istat, Roma.

RUSSO A. (1985), *Modelli per la presentazione degli errori standard in campioni complessi, Giornate di metodologia statistica*, Bressanone.

RUSSO A. e FALORSI P.D. (1985), *Rilevazioni campionarie delle forze di lavoro: Metodologia del campionamento, calcolo e presentazione errori campionari*, Quaderni di Discussione n. 6, Istat, Roma.

RUSSO A. e ZONNO G. (1986), *Indagine sulle opinioni e gli atteggiamenti degli italiani sulle tendenze demografiche: Piano della rilevazione campionaria, riporto dei dati all'universo, calcolo e presentazione errori campionari*, Working Paper, Istituto di Ricerche sulla Popolazione, CNR, Roma.

RUSSO A. (1987), *Sulla presentazione degli errori di campionamento mediante modelli: Il metodo dei modelli regressivi*, Quaderni di Discussione, n. 4, Istat, Roma.

VERMA V. e PEARCE M.C. (1978), *Users Manual for Clusters, World Fertility Survey*, London.

VERMA V., SCOTT C. e O'MUIRCHEARTAIGH (1980), *Sample Designs and Sampling Errors for the World Fertility Survey, Journal of Statistical Society*, A, Part. 4.

VERMA V. (1982), *The Estimation and Presentation of Sampling Errors*, Technical Bulletin n. 7, World Fertility Survey, New York.

ZANNELLA F. (1982), *Indagine sulle condizioni di salute della popolazione italiana e sul ricorso ai servizi sanitari: Calcolo degli errori di campionamento*, Supplemento al Bollettino Mensile di Statistica, n. 12, Istat, Roma.

ZANNELLA F. (1982), *Piano della rilevazione campionaria degli errori di campionamento*, in Indagine sulla fecondità in Italia, Anno 1979, Università di Padova, Firenze, Roma.

RICERCHE STATISTICHE SULLE MISURE DEGLI EFFETTI DEL DISEGNO DI CAMPIONAMENTO

di *Aldo Russo*

1. INTRODUZIONE

Il tema «L'analisi statistica dell'Effetto del disegno di campionamento» vive oggi un momento di grande diffusione e vitalità scientifica, come mostrano gli articoli sempre più numerosi che appaiono sulle riviste internazionali più prestigiose.

L'importanza del tema nasce dal fatto che esso rappresenta un filone di metodologie che trovano sempre più vasta ed articolata applicazione in vari campi di ricerca; in special modo, i teorici del campionamento hanno dato un impulso decisivo al loro sviluppo ed alla loro applicazione nel quadro dei problemi per la realizzazione delle indagini campionarie condotte su larga scala (1).

A questo riguardo vi è da dire che la problematica posta in essere dall'effetto del disegno di campionamento, ancorché affrontata da singoli studiosi, vede da tempo impegnati anche i maggiori centri di informazione statistica a livello internazionale, tra cui il Bureau of Census degli Stati Uniti, lo Statistics Canada, l'Australian Bureau of Statistics.

In Italia la letteratura sulla teoria dell'effetto del disegno di campionamento è scarsa e soltanto nel corso di questi ultimi anni sono stati dati contributi di un certo rilievo (2).

L'Istat ha dedicato un'attenzione e uno spazio via via crescenti al tema, sviluppando filoni di ricerca sia di carattere essenzialmente metodologico, sia di carattere meramente empirico.

L'idea è stata suggerita dalla lettura di un volume di Kish (1965), mentre lo spunto è stato offerto dalla rilettura critica di un articolo di Verma, Scott e O'Muircheartaigh (1980), contenente un poderoso studio finalizzato fondamentalmente alla valutazione dei piani di campionamento utilizzati da vari Paesi per effettuare l'indagine mondiale sulla fecondità.

Tali trattazioni, anche se forniscono un fecondo inquadramento concettuale della teoria dell'effetto del disegno di campionamen-

to, lasciano aperti ed insoluti alcuni aspetti di fondo concernenti i difficili problemi:

— della ricerca di un approccio metodologico unitario;

— della metodologia statistica da seguire per la stima dell'effetto del disegno di campionamento e delle sue principali componenti.

Nel paragrafo successivo offriremo una sintesi di alcuni contributi teorici sviluppati nel nostro Istituto che rappresentano un tentativo di aprire la strada verso il raggiungimento di una doverosa chiarificazione concettuale, logica e metodologica nel senso dei problemi sopra indicati, anche se il traguardo sembra ancora piuttosto lontano.

Verranno poi illustrati i principali risultati delle ricerche empiriche (Par. 3). Infine, alcune considerazioni sui problemi ancora aperti completeranno l'esposizione.

2. CONTRIBUTI METODOLOGICI

2.1 Breve digressione sul significato di deff

Per la realizzazione delle indagini campionarie condotte su larga scala, come sono tipicamente quelle dell'Istat, si ricorre a disegni di campionamento che nella letteratura statistica corrente sono definiti «complessi».

Nella maggior parte dei casi, infatti, si tratta di disegni stratificati, a più stadi di selezione, basati su una struttura probabilistica di estrazione del campione, secondo cui le unità degli stadi di ordine più elevato vengono estratte con probabilità proporzionale all'ampiezza e le unità finali con probabilità uguali senza reimmissione.

Nel processo di ripartizione del campione viene generalmente rispettata la condizione di autoponderazione dei valori campionari per la determinazione di stime centrate [3].

Per l'ottenimento delle stime delle caratteristiche della popolazione oggetto d'indagine si ricorre frequentemente all'impiego di stimatori del rapporto, separato o combinato, con post-stratificazione delle unità finali.

Si tratta, inoltre, di indagini che hanno la finalità di produrre un numero elevato di stime, in genere di tipo diverso (medie, totali, fre-

quenze assolute e relative, rapporti, ecc.); è utile, infine, sottolineare che la dimensione campionaria viene determinata sotto il vincolo che il campione risponda ad obiettivi di efficienza anche per diverse sottoclassi della popolazione oggetto di studio.

Ciascuno degli aspetti caratterizzanti la strategia campionaria sopra illustrata merita una particolare attenzione, in quanto esercita un'azione che si traduce in un effetto sulla varianza di campionamento e quindi sul livello di affidabilità dei risultati forniti dalle indagini stesse.

Per misurare l'effetto globale imputabile a tale costellazione di azioni, i cui effetti di segno diverso presentano in genere diversa importanza relativa, Kish (1965) ha suggerito il coefficiente deff, noto appunto col nome di «Effetto del disegno di campionamento (o Design Effect)», espresso dal rapporto:

$$\text{deff}\,(\hat{Y}) = \frac{V\,(\hat{Y})}{V\,(\hat{Y}_{ccs})} \qquad (1)$$

in cui $V\,(\hat{Y})$ e $V\,(\hat{Y}_{css})$ indicano rispettivamente la varianza della stima $\hat{Y}$ ottenuta con una strategia campionaria complessa, come è quella su descritta, e quella di un (ipotetico) campione casuale semplice di uguale numerosità in termini di unità finali in cui la stima $\hat{Y}_{css}$ si suppone ottenuta mediante uno stimatore semplice.

Nel caso delle indagini Istat condotte sulle famiglie, ad esempio, deff riflette dunque l'effetto collettivo delle azioni esercitate dalla stadificazione, dalla stratificazione delle unità primarie, dalla post-stratificazione delle unità finali, dallo stimatore del rapporto e dalla eventuale introduzione di pesi variabili per l'ottenimento delle stime.

Un passo ulteriore nell'approfondimento dello studio di deff è stato compiuto dallo stesso Kish, relativamente al problema di enucleare e misurare le principali componenti di deff. Anche Verma, Scott e O'Muircheartaigh si sono occupati della medesima questione nell'interessante ricerca già citata, il cui scopo essenziale è la proposizione di un impianto metodologico finalizzato alla determinazione di una stima degli effetti stratificazione, stadificazione e ponderazione.

Per esprimere quantitativamente i suddetti effetti, gli Autori citati hanno suggerito i tre seguenti indici:

$$E_s = \frac{V\,(\hat{Y})}{V\,(\hat{Y}_{\bar{s}})}; \quad E_c = \frac{V\,(\hat{Y})}{V\,(\hat{Y}_{\bar{c}})}; \quad E_p = \frac{V\,(\hat{Y})}{V\,(\hat{Y}_a)} \qquad (2)$$

in cui, con riferimento ad esempio alle indagini Istat sulle famiglie, $V(\hat{Y}_{\bar{s}})$ rappresenta la varianza della stima $\hat{Y}_{\bar{s}}$ ottenuta mediante una strategia campionaria (fittizia) uguale a quella (reale) adottata per l'ottenimento di $\hat{Y}$, ad eccezione del fatto che le unità primarie (comuni) non sono stratificate; $V(\hat{Y}_{\bar{c}})$ indica la varianza della stima $\hat{Y}_{\bar{c}}$ ottenuna con una strategia campionaria (fittizia) costituita da un disegno in cui le unità di campionamento sono le sole famiglie (grappoli di individui), fermi restando il processo di stratificazione e la struttura formale dello stimatore; $V(\hat{Y}_a)$ rappresenta la varianza della stima $\hat{Y}_a$ determinata con una strategia campionaria (fittizia), uguale a quella usata per la determinazione di $\hat{Y}$, che presenta inoltre la proprietà di rispettare la condizione di autoponderazione per la generazione di stime centrate delle caratteristiche della popolazione oggetto di studio.

Nel porre la questione in questi termini il problema della determinazione della misura più idonea delle varianze $V(\hat{Y}_{css})$, $V(\hat{Y}_{\bar{s}})$, $V(\hat{Y}_{\bar{c}})$ e $V(\hat{Y}_a)$ è ben lungi dall'essere risolto. Sorgono, infatti, molti problemi di ordine concettuale e metodologico, ma il problema di fondo, a nostro avviso, è costituito dalle complesse questioni connesse al tipo di metodologia che può seguirsi per la stima delle varianze suddette.

Gli autori sopra citati hanno affrontato tale problema nel contesto di una filosofia del compromesso fra rigore logico e formale, da una parte, e praticità empirica, dall'altra; il loro orientamento concettuale e metodologico, pur presentando senza dubbio un notevole interesse per l'elevata utilità operativa che caratterizza i procedimenti di stima da loro utilizzati, non consente di inquadrare e risolvere in modo completamente razionale e statisticamente soddisfacente il problema della stima degli effetti sopra illustrati.

Le limitazioni teoriche, come vedremo in modo più dettagliato nel seguito, nascono dal fatto che i procedimenti seguiti per la determinazione di una stima delle varianze $V(\hat{Y}_{css})$, $V(\hat{Y}_{\bar{s}})$, $V(\hat{Y}_{\bar{c}})$ e $V(\hat{Y}_a)$, basati necessariamente sull'utilizzazione delle informazioni dell'indagine concreta, trascurano (completamente o in parte) la circostanza che le informazioni stesse provengono da campioni complessi.

Nel successivo paragrafo illustreremo le considerazioni fatte sul piano concettuale e metodologico nel quadro delle ricerche, effettuate nel nostro Istituto, volte a contribuire alla costruzione di una teoria più corretta e soddisfacente in tema di stima degli effetti citati.

2.2. Riflessioni e proposte metodologiche

2.2.1 *Un breve excursus storico*

La produzione scientifica del Reparto Studi dell'Istat, in tema di deff, si colloca naturalmente nel panorama delle indagini del nostro Istituto; tutte le ricerche sono state svolte nel contesto del campionamento a due stadi, che è alla base delle rilevazioni condotte sulle famiglie e di altre rilevazioni, come quelle effettuate sulle aziende agricole, che hanno la finalità di produrre stime della consistenza del patrimonio bovino, suino, ovino e caprino.

Tale scelta non è riduttiva in quanto le metodologie suggerite con riferimento a tali indagini, mediante semplici adattamenti formali, possono essere particolarizzate al caso delle rimanenti indagini basate su disegni campionari e stimatori più semplici.

I primi lavori (Coccia, 1986; Russo, 1985; Russo, 1986a; Russo, 1986b) si riferiscono al caso di indagini campionarie a due stadi con stratificazione delle unità primarie, in cui si ipotizza un solo livello di clusterizzazione; per la selezione del campione si assume che le unità primarie e quelle secondarie vengano scelte con probabilità uguali e senza reimmissione. Per la determinazione delle stime dell'indagine si fa infine l'ipotesi di ricorrere all'utilizzazione dello stimatore diretto basato sulle probabilità di inclusione.

L'obiettivo comune di tali lavori è fornire, con un approccio globale ed unitario, una stima soddisfacente di deff e degli effetti stratificazione, stadificazione e ponderazione.

Un passo ulteriore in questa direzione, seguendo lo stesso filo logico e metodologico, è stato compiuto da Russo (1986c), il quale ha affrontato il problema della stima degli effetti testé menzionati, assumendo una struttura probabilistica di selezione delle unità secondo cui le unità primarie sono estratte con reimmissione e probabilità proporzionale all'ampiezza e quelle secondarie senza reimmissione e probabilità uguali.

Un altro contributo di rilievo è stato dato con una ricerca (Coccia, D'Angiolini, Falorsi e Russo, 1987) il cui scopo essenziale era la proposizione di una metodologia per la stima degli effetti in questione con riferimento all'indagine campionaria sulle forze di lavoro. La formulazione di tale metodologia ha comportato maggiori difficoltà, d'ordine sia concettuale che statistico, in conseguenza del

fatto che per l'ottenimento delle stime dell'indagine si ricorre all'uso di uno stimatore del rapporto separato post-stratificato e che da ogni strato viene scelta una sola unità primaria. In tali circostanze, sorgono tutte le complesse questioni concernenti la determinazione di una espressione esatta della varianza e del calcolo di una stima della stessa.

Una naturale estensione delle precedenti ricerche è rappresentata dagli studi effettuati nell'ambito delle indagini basate su campioni spazio-tempo (Coccia e Russo, 1987; Falorsi, 1987).

Infine, i lavori di Russo (1988) e Falorsi P. e Falorsi S. (1989) sono volti a risolvere il problema della stima degli effetti in esame nel caso di indagini a due stadi, con due livelli di clusterizzazione, in cui le unità primarie sono selezionate con probabilità proporzionale all'ampiezza e senza reimmissione, e quelle secondarie con probabilità uguali e senza reimmissione.

2.2.2 *Aspetti metodologici fondamentali dell'approccio suggerito per lo studio degli effetti del disegno*

Consideriamo la seguente situazione: immaginiamo di aver effettuato un'indagine (che indicheremo d'ora innanzi con il simbolo I) allo scopo di determinare una stima del totale Y, mediante un campione a due stadi con stratificazione delle unità di primo stadio. Facciamo ancora l'ipotesi che per l'ottenimento della stima di Y sia stato adottato uno stimatore diretto e centrato e che il campione non sia autoponderante. Per il momento non introduciamo il meccanismo probabilistico di selezione del campione, che sarà tuttavia definito nel seguito.

Al fine di rendere più chiara e concreta l'esposizione supponiamo che le unità di primo stadio (UP) siano costituite dai comuni e quelle di secondo stadio (US) dagli individui residenti nei comuni stessi.

Indichiamo inoltre con $\hat{Y}$ la stima del totale Y e con $\hat{V}(\hat{Y})$ la stima della varianza campionaria $V(\hat{Y})$.

Per una maggiore comprensione dei successivi sviluppi è utile introdurre una simbologia per descrivere le caratteristiche strutturali della popolazione e del campione dell'indagine I.

Indichiamo con:

i	indice di UP
j	indice di US
h	indice di strato (h = 1,...H)
N_h	numero di UP nello strato h
N	numero complessivo di UP
M_{hi}	numero di US nell'UP (hi)
M_h	numero di US nello strato h
M	numero totale di US
n_h	numero di UP-campione nello strato h
n	numero totale di UP-campione
m_{hi}	numero di US-campione nell'UP-campione (hi)
m_h	numero di US-campione nello strato h
m	numero totale di US-campione

Consideriamo ora le quattro strategie campionarie «fittizie» costituite dai seguenti disegni campionari:

D_1: a due stadi semplici (non stratificato) uguale a quello adottato per l'indagine I;

D_2: ad uno stadio stratificato, in cui le unità di campionamento sono costituite da soli individui e la dimensione campionaria del generico strato h (h = 1,...H) è pari a $m_h = \sum_i m_{hi}$, con i = 1,...n_h;

D_3: identico a quello usato per l'indagine I, ma autoponderante; indicando con ${}_am_{hi}$ il numero di US-campione nell'UP-campione (hi) nella ipotesi di autoponderazione si avrà pertanto:
$\sum_h \sum_i m_{hi} = \sum_h \sum_i {}_am_{hi} = m$, con i = 1,..., n_h e h = 1,...,H;

D_4: campione casuale semplice di soli individui, di numerosità complessiva uguale a m.

Associamo ora a ciascuno dei disegni sopra descritti uno stimatore diretto e centrato per la determinazione di una stima del totale Y.

In tal modo restano definite quattro strategie campionarie fittizie in base alle quali è possibile calcolare le sime $\hat{Y}_{\bar{s}}$, $\hat{Y}_{\bar{c}}$, $\hat{Y}_a$ e $\hat{Y}_{ccs}$, a cui corrispondono rispettivamente le varianze V ($\hat{Y}_{\bar{s}}$), V ($\hat{Y}_{\bar{c}}$), V ($\hat{Y}_a$) e V ($\hat{Y}_{ccs}$), coinvolte nelle (1) e (2).

A questo punto nasce il problema cardine costituito dalle complesse questioni connesse alla determinazione di una stima degli effetti espressi dalla (1) e dalla (2).

L'ottenimento di tali stime si presenta notevolmente difficoltoso — sia dal lato concettuale che dal lato della metodologia statistica — in quanto le espressioni delle varianze si riferiscono ai disegni fittizi D_1, D_2, D_3 e D_4, mentre la costruzione delle stime deve essere necessariamente basata sull'utilizzazione delle informazioni desumibili dall'indagine I, che provengono invece da un diverso disegno campionario.

L'approccio suggerito da Kish, Verma ed altri, per la stima della varianza $V(\hat{Y}_{\bar{s}})$, $V(\hat{Y}_{\bar{c}})$, $V(\hat{Y}_a)$ e $V(\hat{Y}_{css})$, poggia su alcune ipotesi che consistono:

a) per il disegno D_1, nell'assumere che gli n comuni-campione dell'indagine I costituiscano un campione casuale semplice estratto da una popolazione di N comuni;

b) per il disegno D_2, nell'assumere che gli m_h ($h = 1, ..., H$) individui costituiscano un campione casuale semplice estratto da una sub-popolazione di M_h unità;

c) nel sostituire al disegno D_3 un disegno ad uno stadio stratificato in cui da ogni strato si suppone estratto un campione casuale semplice di m_h ($h = 1, ..., H$) individui; viene, altresì, ipotizzato che la varianza del generico carattere oggetto di studio sia costante ($S_h^2 = \text{cost}$ per ogni $h \in H$);

d) per il disegno D_4, nell'assumere che gli m individui campione dell'indagine I costituiscano un campione casuale semplice estratto da una popolazione di M unità.

In tale approccio appaiono evidenti almeno quattro incongruenze. La prima si riferisce al fatto che l'ipotesi a) trascura la circostanza che gli n comuni-campione dell'indagine I sono il risultato di un processo di selezione basato sulla ripartizione degli N comuni in H strati. La seconda nasce dalla considerazione che l'ipotesi b) ignora la circostanza che gli m_h individui, relativi al generico strato h del disegno D_2, sono il risultato di un processo di selezione a due stadi: estrazione di n_h comuni campione e successiva estrazione delle US $m_{h1}, ..., m_{hn_h}$ rispettivamente dalle sub-popolazioni $M_{h1}, ..., M_{hn_h}$. La terza concerne: i) la sostituzione del disegno D_3 (o più in generale di disegni più complessi di D_3) con un disegno ad uno stadio stra

tificato; ii) l'ipotesi S_h^2 = cost, che in generale non è vera. La quarta, infine, scaturisce dalla considerazione che il disegno ignora completamente che gli m individui campione derivano da un campione complesso.

È appunto da queste ipotesi, e dalle conseguenti limitazioni teoriche che esse introducono nell'approccio proposto da Kish ed altri, che prende l'avvio e si svolge gran parte delle nostre ricerche, il cui obiettivo comune è quello di tracciare uno schema metodologico che, basandosi sulla rimozione delle suddette ipotesi, consenta di affrontare in modo più soddisfacente il problema di ricavare una misura delle stime degli effetti del disegno.

Nelle pagine che seguono, illustriamo le linee metodologiche dell'approccio da noi suggerito, senza tuttavia approfondire i dettagli tecnici dei vari metodi di stima, peraltro ampiamente illustrati nei lavori precedentemente citati.

Per dare ordine alla nostra esposizione cominciamo col considerare la situazione relativamente più semplice, anche se non riveste una notevole importanza nel quadro della realizzazione delle indagini campionarie effettive: cioè, con riferimento all'indagine I, supponiamo che il processo di selezione del campione consista nella estrazione sia delle UP che delle US con probabilità uguali e senza reimmissione.

Per esporre la metodologia in questione definiamo in primo luogo l'espressione dello stimatore del totale Y, con riferimento all'indagine I.

Nel campionamento a due stadi, con stratificazione delle UP, in cui le unità sono estratte secondo il meccanismo probabilistico su descritto, si dimostra che una stima corretta di Y è fornita dallo stimatore:

$$\hat{Y} = \sum_h \sum_i \sum_j K_{hij}\, Y_{hij} \qquad (3)$$

in cui: $h = 1, ..., H$; $i = 1, ..., n_h$, $j = 1, ..., m_{hi}$; Y_{hij} = valore del carattere y osservato sull'US j dell'UP i dello strato h; K_{hij} = peso attribuito all'US (hij) [4].

Ai fini degli sviluppi successivi conviene introdurre anche le seguenti forme equivalenti di (3):

$$\hat{Y} = \sum_h \hat{Y}_h = \sum_h \sum_i \frac{N_h}{n_h} \hat{Y}_{hi} \tag{4}$$

in cui:

$$\hat{Y}_{hi} = \sum_j \frac{M_{hi}}{m_{hi}} Y_{hij} \tag{5}$$

La varianza di $\hat{Y}$ è definita a sua volta dalla relazione:

$$V(\hat{Y}) = \sum_h \left[\frac{N_h (N_h - n_h)}{n_h} {}_aS_h^2 + \frac{N_h}{n_h} \sum_i \frac{M_{hi} (M_{hi} - m_{hi})}{m_{hi}} {}_bS_{hi}^2 \right] \tag{6}$$

in cui:

$${}_aS_h^2 = \frac{1}{N_h - 1} \sum_j (Y_{hi} - \bar{Y}_h)^2 \tag{7}$$

$${}_bS_{hi}^2 = \frac{1}{M_{hi} - 1} \sum_j (Y_{hi} - \bar{\bar{Y}}_{hi})^2 \tag{8}$$

$$Y_{hi} = \sum_j Y_{hij}\ ;\ Y_h = \sum_i Y_{hi}\ ;\ \bar{Y}_h = \frac{Y_h}{N_h}\ ;\ \bar{\bar{Y}}_{hi} = \frac{Y_{hi}}{M_{hi}} \tag{9}$$

con $i = 1, ..., N_h$ e $j = 1, ..., M_{hi}$.

Infine, è possibile mostrare che una stima corretta di $V(\hat{Y})$ si ottiene traducendo in termini campionari la (6), cioè:

$$\hat{V}(\hat{Y}) = \sum_h \left[\frac{N_h (N_h - n_h)}{n_h} {}_as_h^2 + \frac{N_h}{n_h} \sum_i \frac{M_{hi} (M_{hi} - m_{hi})}{m_{hi}} {}_bs_{hi}^2 \right] \tag{10}$$

in cui:

$${}_as_h^2 = \frac{1}{n_h - 1} \sum_i (\hat{Y}_{hi} - \hat{\bar{Y}}_h)^2 \tag{11}$$

$${}_bs_{hi}^2 = \frac{1}{m_{hi} - 1} \sum_j (Y_{hij} - \hat{\bar{\bar{Y}}}_{hi})^2 \tag{12}$$

$$\hat{\bar{Y}}_h = \frac{1}{n_h} \hat{Y}_h;\ \hat{\bar{\bar{Y}}}_{hi} = \frac{1}{m_{hi}} \sum_j Y_{hij} \tag{13}$$

con $i = 1, ..., n_h$ e $j = 1, ..., m_{hi}$.

Ciò premesso, affrontiamo ora il problema di una stima dell'effetto stratificazione, E_s, mediante il quale è possibile quantificare il guadagno nel livello di precisione dovuto all'azione esercitata dal processo di stratificazione delle UP.

A tale scopo consideriamo la strategia campionaria fittizia costituita dal disegno D_1 e dallo stimatore diretto e centrato. Potremo individuare strutture e caratteristiche di tale strategia mediante una particolarizzazione della simbologia e delle espressioni definite per il campione a due stadi con stratificazione delle UP.

Segue pertanto che una stima corretta del totale Y, ricorrendo ad un campione a due stadi semplici e a uno stimatore diretto e centrato, è fornita dall'epressione [5]:

$$\hat{Y}_{\bar{s}} = \sum_i \sum_j K_{ij} Y_{ij} \tag{14}$$

con $i = 1, ..., n$ e $j = 1, ..., m_i$.

Dalla (6) si ottiene poi la corrispondente espressione della varianza:

$$V(\hat{Y}_{\bar{s}}) = \frac{N(N - n)}{n} {}_aS^2 + \frac{N}{n} \sum_i \frac{M_i (M_i - m_i)}{m_i} {}_bS_i^2 \tag{15}$$

in cui:

$${}_aS^2 = \frac{1}{N - 1} \sum_i (Y_i - \bar{Y})^2 \tag{16}$$

$${}_bS_i^2 = \frac{1}{M_i - 1} \sum_j (Y_{ij} - \bar{\bar{Y}}_i)^2 \tag{17}$$

$$Y_i = \sum_j Y_{ij} \; ; \; Y = \sum_i Y_i \; ; \; \bar{Y} = \frac{Y}{N} \; ; \; \bar{\bar{Y}}_i = \frac{Y_i}{M_i} \tag{18}$$

A questo punto, seguendo l'approccio suggerito da Kish ed altri, per la determinazione di una stima di $V(\hat{Y}_{\bar{s}})$ occorre usare un'espressione avente una struttura analoga alla (10), eliminando però l'indice h di strato; procedendo in tal modo, come abbiamo già sottolineato, si trascura però la circostanza che l'insieme delle n UP è costituito da n_1 unità estratte dalle N_1 appartenenti allo strato 1, da n_2 estratte dalle N_2 UP dello strato 2, e così via.

Il metodo da noi seguito si propone invece di trovare una stima corretta di V ($\hat{Y}_{\bar{s}}$) nel rispetto della condizione che le n UP provengono da un universo stratificato, risultando così concettualmente più adeguato alla realtà e statisticamente più razionale e soddisfacente.

Il metodo parte dal considerare un'espressione, equivalente alla formula (15), ottenuta esprimendo quest'ultima col simbolismo del campione a due stadi stratificato, ossia:

$$V(\hat{Y}_{\bar{s}}) = W \left[\sum_h (N_h - 1)\, {}_aS_h^2 + \sum_h \frac{Y_h^2}{N_h} - \frac{Y^2}{N} \right] + \tag{19}$$

$$+ \sum_h \sum_i W_{hi}\, {}_bS_{hi}^2$$

$$\text{in cui } W = \frac{N(N-n)}{n(N-1)} \; ; \; W_{hi} = \frac{N\, M_{hi}(M_{hi} - m_{hi})}{n\, m_{hi}} \tag{20}$$

La relazione (19) è strutturata in modo tale da consentire l'ottenimento di una stima corretta di V ($\hat{Y}_{\bar{s}}$): a tal fine è sufficiente determinare una stima corretta di ${}_aS_h^2$, Y_h^2 e Y^2 e dell'ultimo addendo della (19).

Nel lavoro di Russo (1985), attraverso derivazioni algebriche descritte in modo dettagliato, abbiamo mostrato che le seguenti espressioni:

$${}_as_h^2 - \frac{1}{n_h} \sum_i \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}}\, {}_bs_{hi}^2 \qquad (i = 1, \ldots, n_h) \tag{21}$$

$$\hat{Y}_h^2 - \hat{V}(\hat{Y}_h) \tag{22}$$

$$\hat{Y}^2 - \hat{V}(\hat{Y}) \tag{23}$$

$$\sum_h \sum_i \frac{N_h}{n_h} W_{hi}\, {}_bs_{hi}^2 \qquad (i = 1, \ldots, n_h) \tag{24}$$

forniscono rispettivamente una stima corretta di ${}_aS_h^2$, Y_h^2, Y^2 e dell'ultimo addendo della (19).

Introducendo le (21), (22), (23) e (24) nella (19) si ottiene che una stima corretta di $V(\hat{Y}_{\bar{s}})$ è data da:

$$\hat{V}(\hat{Y}_{\bar{s}}) = W\left[\sum_h (N_h - 1)\, {}_a s_h^2 + \sum_h \frac{\hat{Y}_h^2}{N_h} - \frac{\hat{Y}^2}{N} + \frac{\hat{V}(\hat{Y})}{N} - \sum_h \frac{\hat{V}(\hat{Y}_h)}{N_h}\right] +$$

$$+ \sum_h \frac{N(N-n) + N_h(n-1)}{nn_h(N-1)} \sum_i \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}}\, {}_b s_{hi}^2 \qquad (25)$$

Una volta determinata $\hat{V}(\hat{Y}_{\bar{s}})$, si può ottenere una misura dell'efficacia della stratificazione delle UP attraverso il calcolo di E_s, definito dal primo rapporto figurante nelle (2).

Passiamo ora ad affrontare il problema di ottenere una stima dell'effetto stadificazione, che consente di quantificare l'aumento di varianza imputabile all'introduzione dei comuni come primo stadio di campionamento.

Tenendo presenti le considerazioni già svolte, per la soluzione di tale problema basta determinare una stima della varianza $V(\hat{Y}_{\bar{c}})$.

A tale scopo faremo pertanto riferimento ad una strategia costituita del disegno campionario D_2 e dallo stimatore diretto e centrato.

Utilizzando il simbolismo adottato per il campione dell'indagine I, salvo l'eliminazione dell'indice i di UP, richiamiamo le espressioni della stima del totale Y e della corrispondente varianza di campionamento relativi alla strategia testé menzionata.

Per il campionamento ad uno stadio stratificato in cui le unità sono estratte con probabilità uguali e senza remmissione, è possibile mostrare che una stima corretta del totale Y è fornita dallo stimatore [6]:

$$\hat{Y}_{\bar{c}} = \sum_h \sum_j K_{hj}\, Y_{hj} \qquad (26)$$

in cui: Y_{hj} indica il valore del carattere y osservato sull'US (hj); K_{hj} rappresenta il peso attribuito all'unità stessa; $j = 1, \ldots, m_h$ e $h = 1, \ldots, H$.

La varianza di $\hat{Y}_{\bar{c}}$ è fornita dall'espressione:

$$V(\hat{Y}_{\bar{c}}) = \sum_h \frac{M_h(M_h - m_h)}{m_h} S_h^2 \qquad (27)$$

in cui:

$$S_h^2 = \frac{1}{M_h - 1} \sum_j (Y_{hj} - \bar{\bar{Y}}_h)^2 \qquad (28)$$

$$Y_h = \sum_j Y_{hj} \; ; \; \bar{\bar{Y}}_h = \frac{Y_h}{M_h} \qquad (29)$$

nelle quali $j = 1, ..., M_h$.

Ciò premesso, e prima di illustrare il nostro metodo per ottenere una stima di $V(\hat{\bar{Y}}_c)$, riteniamo opportuno descrivere quello proposto da Kish ed altri.

Il metodo seguito dai citati Autori si basa sull'utilizzazione della nota espressione:

$$\hat{V}(\hat{\bar{Y}}_c) = \sum_h \frac{M_h (M_h - m_h)}{m_h} s_h^2 \qquad (30)$$

in cui:

$$s_h^2 = \frac{1}{m_h - 1} \sum_j (Y_{hj} - \hat{\bar{Y}}_h)^2; \; \hat{\bar{Y}}_h \frac{1}{m_h} \sum_j Y_{hj} \qquad (31)$$

in cui $j = 1, ..., m_h$.

La limitazione teorica del procedimento in questione nasce dal fatto che esso ignora la circostanza che le m_h US provengono da un processo di selezione a due stadi.

Nel lavoro sopra citato (Russo, 1985) abbiamo indicato un procedimento che tiene conto di quest'ultima circostanza e che si propone di ottenere una stima corretta di $V(\hat{\bar{Y}}_c)$.

Il procedimento parte dal considerare un'espressione equivalente alla (27), ottenuta introducendo in essa il simbolismo del campionamento a due stadi con stratificazione delle UP. L'espressione in esame è definita dalla seguente struttura formale:

$$V(\hat{\bar{Y}}_c) = \sum_h W_h \left[\sum_i (M_{hi} - 1) \, {}_bS_{hi}^2 + \sum_i \frac{Y_{hi}^2}{M_{hi}} - \frac{Y_h^2}{M_h} \right] \qquad (32)$$

in cui:

$$W_h = \frac{M_h (M_h - m_h)}{m_h (M_h - 1)} \qquad \text{con } i = 1, ..., N_h.$$

Per ricavare una stima corretta della (32) è sufficiente ottenere una stima corretta dei tre addendi compresi fra le parentesi quadre.

Nell'articolo citato abbiamo mostrato che una stima corretta della varianza $V(\hat{Y}_{\bar{c}})$ è fornita dall'espressione:

$$\hat{V}(\hat{Y}_{\bar{c}}) = \sum_h W_h \left[\frac{N_h}{n_h} \sum_i \frac{M_{hi}}{m_{hi}} \sum_j Y_{hij}^2 - \frac{\hat{Y}_h^2}{M_h} + \frac{\hat{V}(\hat{Y}_h)}{M_h} \right] \qquad (33)$$

A questo punto, avendo ottenuto il risultato cercato, si può determinare una stima dell'effetto E_c.

Occupiamoci ora del problema relativo alla determinazione di una stima dell'effetto ponderazione, E_p.

L'importanza del calcolo di tale stima nasce dal fatto che, in generale, la varianza di un campione non autoponderante (come è quello adottato per l'indagine I) è non inferiore a quella ottenibile con un campione uguale a quello usato per l'indagine I, ma autoponderante.

L'aumento di varianza, determinato dall'introduzione di pesi variabili, per l'ottenimento della stima $\hat{Y}$, è appunto misurabile in termini di E_p. Per ottenere una stima di E_p, avendo già calcolato $\hat{V}(\hat{Y})$, basterà ricavare una stima di $V(\hat{Y}_a)$.

Il metodo escogitato da Kish, basato sulle ipotesi precedentemente illustrate, fornisce per la stima dell'effetto ponderazione l'espressione seguente:

$$\hat{E}_p = \frac{(\sum_h m_h)(\sum_h m_h \tilde{K}_h^2)}{(\sum m_h \tilde{K}_h)^2} \qquad (34)$$

in cui: $\tilde{K}_h = M_h / m_h$.

Dalla (34) si evince che $\hat{E}_p$ è un fattore costante per tutte le variabili: cioè per ciascuna delle stime oggetto di indagine è costante l'incidenza — sulle corrispondenti varianze campionarie — dovuta all'introduzione di pesi variabili per la determinazione delle stime medesime.

Il criterio da noi proposto, basato sulla strategia campionaria costituita dal disegno D_3 e dallo stimatore diretto e centrato, si propone di trovare una stima corretta di $V(\hat{Y}_a)$, che tiene conto della circostanza che le informazioni desumibili dall'indagine I sono il risultato di un processo campionario non autoponderante.

A tale scopo indichiamo con:

$$\hat{Y}_a = K \sum_h \sum_i \sum_j Y_{hij} \tag{35}$$

in cui:

$$K = \frac{N_h}{n_h} \frac{M_{hi}}{{}_a m_{hi}} \tag{36}$$

la stima corretta del totale Y che avremmo ottenuto se per l'indagine l fosse stata adottata una strategia campionaria autoponderante.

Mediante semplici passaggi (Russo, 1986a), partendo dalla (69), si può mostrare che la varianza di $\hat{Y}_a$ può porsi nella forma:

$$V(\hat{Y}_a) = \sum_h \left[\frac{N_h (N_h - n_h)}{n_h} {}_a S_h^2 + (K - \frac{N_h}{n_h}) \sum_i M_{hi} {}_b S_{hi}^2 \right] \tag{37}$$

Nel lavoro appena citato abbiamo determinato un'espressione che fornisce una stima corretta di $V(\hat{Y}_a)$, che riteniamo utile riportare:

$$\hat{V}(\hat{Y}_a) = \sum_h \frac{N_h (N_h - n_h)}{n_h} {}_a s_h^2 +$$

$$+ \frac{N_h}{n_h} \sum_i M_{hi} \left(\frac{M_{hi}}{{}_a m_{hi}} - \frac{M_{hi}}{m_{hi}} \right) {}_b s^2{}_{hi} \tag{38}$$

in cui $h = 1, ..., H$ e $i = 1, ..., n_h$.

Il calcolo della (38) richiede la determinazione delle sole n numerosità ${}_a m_{hi}$, essendo noti i valori di N_h, n_h, M_{hi}, ${}_a s_h^2$ e ${}_b s_{hi}^2$.

Il calcolo delle dimensioni ${}_a m_{hi}$ può effettuarsi mediante l'uso della relazione:

$${}_a m_{hi} = \frac{N_h M_{hi}}{n_h K} \tag{39}$$

con K definito da:

$$K = \frac{\sum_h \frac{N_h}{n_h} \sum_i M_{hi}}{m} \qquad (i = 1, ..., n_h) \tag{40}$$

Passiamo, infine, ad esporre le linee metodologiche fondamentali del criterio indicato per la stima di deff.

A tal fine, con riferimento alla strategia costituita dal disegno campionario D_4 e dallo stimatore diretto e centrato, definiamo le espressioni della stima $\hat{Y}_{ccs}$, del totale Y, e della corrispondente varianza campionaria.

Si ha:

$$\hat{Y}_{ccs} = \frac{M}{m} \sum_j Y_j \qquad (j = 1, ..., m) \tag{41}$$

$$V(\hat{Y}_{ccs}) = \frac{M(M - m)}{m} S^2 \tag{42}$$

in cui:

$$S^2 = \frac{1}{M - 1} \sum_j (Y_j - \bar{\bar{Y}})^2; \qquad \bar{\bar{Y}} = \frac{Y}{M} \qquad (j = 1, ..., M) \tag{43}$$

Il criterio suggerito da Kish per il calcolo di una stima della (42) si basa sull'utilizzazione della nota espressione:

$$\hat{V}(\hat{Y}_{ccs}) = \frac{M(M - m)}{m} s^2 \tag{44}$$

in cui:

$$s^2 = \frac{1}{m - 1} \sum_j (Y_j - \hat{\bar{Y}})^2; \qquad \hat{\bar{Y}} = \frac{1}{m} \sum_j Y_j \tag{45}$$

che, come abbiamo già osservato, si fonda sull'assunzione semplificatrice che le m unità rappresentino un campione casuale semplice estratto da una popolazione di M unità.

Per superare tale limitazione osserviamo anzitutto che la (42) può essere efficacemente espressa dalla relazione seguente:

$$V(\hat{Y}_{ccs}) = W^1 \left[\sum_h \sum_i (M_{hi} - 1)\, {}_bS^2_{hi} + \right.$$

$$\left. + \sum_h \sum_i \frac{Y^2_{hi}}{M_{hi}} - \frac{Y^2}{M} \right] \tag{46}$$

in cui:

$$W^1 = \frac{M\,(M - m)}{m\,(M - 1)} \qquad (47)$$

L'opportunità di tradurre la (42) nella (46), col simbolismo introdotto con riferimento all'indagine I, poggia ancora sul principio — alla base di tutto il nostro impianto metodologico — che la metodologia di stima di $V(\hat{Y}_{ccs})$ deve tener conto del fatto che le sole informazioni utilizzabili sono quelle desumibili dall'indagine I, e che esse sono il risultato di un processo campionario avente caratteristiche diverse da quelle di D_4.

Dalla (46) è possibile derivare (Russo, 1985) l'espressione seguente:

$$\hat{V}(\hat{Y}_{ccs}) = W^1 \left[\sum_h \frac{N_h}{n_h} \sum_i \frac{M_{hi}}{m_{hi}} \sum_j Y^2_{hij} - \frac{\hat{Y}^2}{M} + \frac{\hat{V}(\hat{Y})}{M} \right] \qquad (48)$$

che fornisce una stima corretta di $V(\hat{Y}_{ccs})$, mediante la quale è possibile ottenere una stima di deff.

2.2.3. Ulteriori considerazioni sullo studio di deff

Nel paragrafo precedente abbiamo illustrato gli aspetti essenziali della metodologia per la determinazione di una stima di deff e delle sue principali componenti, E_s, E_c e E_p.

Tale metodologia è stata sviluppata con riferimento al caso di indagini basate su una strategia costituita da:

— un disegno campionario a due stadi con stratificazione delle UP, in cui sia le UP che le US sono estratte con probabilità uguali e senza reimmissione;

— uno stimatore diretto e centrato, per l'ottenimento delle stime delle caratteristiche della popolazione oggetto di indagine.

In realtà, per l'effettuazione delle indagini campionarie su larga scala, si ricorre all'impiego di strategia campionarie più complesse capaci di fornire stime più efficienti di quelle ottenibili con la strategia sopra descritta.

A tale scopo è stata da tempo rilevata l'opportunità per i piani di campionamento a due stadi, di estrarre senza sostituzione il cam-

pione delle UP in modo che le loro probabilità di appartenere al campione siano proporzionali alle rispettive misure di un carattere, che possa ritenersi correlato con la caratteristica rispetto alla quale si desidera stimare la popolazione totale. Fin dal 1943 infatti è stato mostrato da Hansen ed Hurwitz che, per il campionamento a due stadi con stratificazione delle UP, la scelta di una sola UP da ciascuno strato con probabilità proporzionale alla sua dimensione è, in generale, più efficiente della scelta effettuata con probabilità uguali.

Viva attenzione è stata inoltre prestata ai problemi riguardanti la costruzione di stimatori che consentano di ottenere stime quanto più efficienti possibile. A questo riguardo, osserviamo che nelle indagini condotte dai maggiori centri di informazioni statistiche a livello internazionale vengono generalmente utilizzati stimatori del rapporto post-stratificati o stimatori composti, espressi comunque come combinazione di stimatori post-stratificati.

In Italia, il disegno standard adottato per le indagini sulle famiglie è a due stadi, con estrazione, da ogni strato, di una sola UP con probabilità di selezione proporzionale alla sua dimensione demografica; per l'ottenimento dei risultati delle indagini, come è stato già sottolineato nei precedenti capitoli, si ricorre all'impiego dello stimatore del rapporto, separato o combinato, con post-stratificazione per sesso e classi di età. Questo disegno campionario presenta inoltre, rispetto a quello preso a base della metodologia precedentemente illustrata, un ulteriore livello di complessità dovuto all'introduzione delle famiglie come secondo stadio di campionamento.

In tali circostanze, lo studio degli effetti del disegno incontra maggiori difficoltà d'ordine sia concettuale che statistico; inoltre, in conseguenza dei nuovi livelli di complessità introdotti dalla strategia campionaria in questione, sorgono altri effetti dovuti all'introduzione delle famiglie come unità di campionamento e della post-stratificazione, E ($\hat{Y}_{ps}$).

Nei lavori citati nel precedente paragrafo 2.2.1, abbiamo suggerito, relativamente a questa strategia campionaria, una metodologia sia per la stima di deff, E_s, E_c e E_p che per la stima dell'effetto dovuto alla post-stratificazione. Lo studio di questi effetti è stato sviluppato seguendo lo stesso filo logico e metodologico alla base dell'approccio già descritto. È da osservare, tuttavia, che per l'introduzione dello stimatore rapporto i metodi proposti conducono a stime

distorte, benché consistenti, delle varianze V ($\hat{Y}_{\bar{s}}$), V ($\hat{Y}_{\bar{c}}$), V ($\hat{Y}_a$), V ($\hat{Y}_{ps}$) e V ($\hat{Y}$).

Riteniamo, infine, opportuno precisare che le espressioni di tali varianze e delle corrispondenti stime sono state derivate linearizzando le stime $\hat{Y}_{\bar{s}}$, $\hat{Y}_{\bar{c}}$, $\hat{Y}_a$, $\hat{Y}_{ps}$ e $\hat{Y}$, mediante lo sviluppo in serie di Taylor arrestato ai termini di ordine lineare.

3. PRINCIPALI RISULTATI DELLE RICERCHE EMPIRICHE

Le prime ricerche (Napolitano, Russo e Zannella, 1983; Russo, 1984) effettuate nel nostro Istituto, che avevano uno scopo essenzialmente descrittivo ed empirico, sono state condotte sull'«Indagine sulle condizioni di salute, anno 1980» e sull'«Indagine sulle vacanze e gli sports, anno 1982».

Per il calcolo degli effetti E_s, E_c, E_p e deff è stato usato l'impianto metodologico proposto da Verma, Scott e O'Muircheartaigh nello studio già citato.

Il calcolo delle varianze campionarie è stato effettuato mediante il programma CLUSTERS, descritto nella relazione di Coccia, contenuta in questo Volume.

I risultati dei calcoli sono riportati nelle tabelle seguenti [7]:

Tabella 1 - Indagine sulle condizioni di salute, anno 1980

Effetti del disegno (*)
(Valori medi per gruppi di variabili)

Gruppi variabili	Campione totale				Dominio non auto-rappresentativo (**)		
	deff	E_s	E_c	E_p	deff	E_s	E_c
1. Socio - demografiche	1,46	0,991	1,09	1,23	1,66	0,983	1,19
2. Malattie in atto	1,51	0,999	1,16	1,23	1,75	0,997	1,29
3. Invalidità permanente	1,39	0,999	1,01	1,23	1,50	0,999	1,02
4. Ricorso ai servizi sanitari	1,53	0,998	1,20	1,23	1,88	0,997	1,39
5. Abitudini al fumo	1,41	0,999	1,14	1,23	1,65	0,998	1,27
Totale variabili	**1,47**	**0,997**	**1,12**	**1,23**	**1,70**	**0,995**	**1,24**

(*) Il calcolo di E_s è stato effettuato ipotizzando un campionamento senza stratificazione per zona altimetrica ed attività economica prevalente dei comuni appartenenti al dominio non autorappresentativo. - (**) È costituito dai comuni con popolazione inferiore a 20.000 abitanti.

Tabella 2 - Indagini sulle vacanze e gli sport, anno 1982.

A: Effetti del disegno: deff., E_s e E_c (*)
(Valori medi per gruppi di variabili)

Gruppi variabili	Campione totale			Dominio non auto-rappresentativo (**)		
	deff	E_s	E_c	deff	E_s	E_c
1. Socio-demografiche	1,62	0,91	1,27	1,73	0,90	1,38
2. Vacanze	2,10	0,91	1,38	2,36	0,90	1,55
3. Sports	1,84	0,98	1,29	2,05	0,98	1,46
4. Giornate di vacanza	2,55	0,97	1,29	3,02	0,97	1,55
Totale variabili	**2,01**	**0,94**	**1,32**	**2,27**	**0,93**	**1,49**

B: Effetto ponderazione entro le regioni

Piemonte	1,22	Friuli-V.G.	1,15	Marche	1,10	Puglia	1,17
Valle d'Aosta	1,15	Liguria	1,17	Lazio	1,37	Basilicata	1,12
Lombardia	1,29	Emilia-R.	1,23	Abruzzo	1,11	Calabria	1,13
Trentino-A.A.	1,09	Toscana	1,31	Molise	1,05	Sicilia	1,13
Veneto	1,13	Umbria	1,17	Campania	1,28	Sardegna	1,13

(*) Il calcolo di E_s è stato effettuato trascurando la stratificazione per classi di ampiezza demografica dei comuni del dominio non autorappresentativo. - (**) È costituito dai comuni con popolazione inferiore a 50.000 abitanti.

Gli aspetti più significativi, che emergono dalle due tabelle, possono essere così riassunti:

— Dall'esame dei valori di E_S, contenuti nella tabella 1, appare evidente che la stratificazione ha un effetto trascurabile sugli errori di campionamento: infatti, la riduzione che si realizza nel campione totale è compresa tra il 9‰ e l'1‰.

Guadagni maggiori, ma comunque modesti, si osservano nella Tab. 2.A: l'errore campionario è in media del 6% inferiore rispetto a quello relativo ad un campione privo di stratificazione.

— Per quanto riguarda l'effetto E_c, dalla tab. 1, si osserva che l'introduzione dei comuni come primo stadio di campionamento, comporta un aumento medio degli errori campionari del 23,8% nei comuni fino a 20.000 abitanti e del 12,2% nel campione totale.

Per l'indagine sulle vacanze e gli sports si riscontrano valori di E_c più elevati: l'aumento medio degli errori campionari è infatti del 49,3% nei comuni fino a 50.000 abitanti e del 31,7 nel campione totale.

Queste percentuali variano sensibilmente con il tipo di variabili considerate. Infatti, da elaborazioni effettuate i cui risultati non so-

no riportati nella presente nota per ragioni di spazio, risulta che per il dominio relativo ai comuni fino a 20.000 abitanti si ha un aumento minimo (1,6%) per le invalidità permanenti e un massimo (39,1%) per il ricorso ai servizi sanitari. Per il campione totale gli effetti risultano, ovviamente più attenuati: 0,9% per le invalidità e 19,5% per il ricorso ai servizi sanitari.

— Infine, per l'effetto ponderazione, dalla tab. 1 si desume che nell'indagine sulle condizioni di salute E_p è uguale a 1,23, per cui l'incremento di errore sofferto a causa dei pesi è del 23% superiore a quello di un campione completamente autoponderante, nel senso che la probabilità di selezione di ciascuna famiglia è ipotizzata la stessa indipendentemente dal comune, dallo strato e dalla regione a cui la famiglia stessa appartiene.

Passando, infine, all'indagine sulle vacanze e gli sports osserviamo anzitutto che gli effetti sono stati calcolati a livello di regione geografica; pertanto i valori di E_p misurano di quanto l'allontanamento dall'autoponderazione incide sulla precisione delle stime regionali. Dall'esame della tab. 2.B emerge che esistono differenze sostanziali tra le regioni: l'incremento minimo (5%) si registra per il Molise, quello massimo (37%) per il Lazio.

4. CONSIDERAZIONI FINALI

Nella presente relazione sono stati illustrati gli aspetti teorici fondamentali ed i principali risultati empirici delle ricerche effettuate in tema di effetti del disegno di campionamento.

Attraverso le ricerche teoriche abbiamo suggerito un insieme di metodologie che, a nostro parere, consentono — nell'ambito di un'ottica unitaria — di risolvere in modo più razionale e statisticamente soddisfacente il problema della stima degli effetti del disegno.

Nel corso degli studi si è avuto occasione di riflettere su alcune complesse questioni che sorgono nel quadro dei problemi volti alla costruzione di strategie campionarie da adottare per l'effettuazione di indagini su larga scala.

Una prima questione di fondo nasce nell'ambito dell'importante problema volto alla determinazione di una funzione atta ad esprimere, in modo statisticamente più soddisfacente, la relazione tra deff ed il coefficiente di omogeneità roh.

A questo riguardo, osserviamo che con riferimento a disegni campionari a due stadi caratterizzati dai seguenti aspetti:

— la UP sono di uguale ampiezza, M;

— da ciascuna UP campione viene selezionato un numero, m, costante di US;

— le UP e le US sono estratte senza reimmissione e probabilità uguali,

è stata derivata, introducendo piccolissime approssimazioni, la relazione seguente:

$$deff = 1 + \varrho\,(m - 1) \qquad (49)$$

in cui ϱ è il coefficiente di correlazione intraclassi, che esprime il grado di similarità fra le US dentro le UP.

Anche Kish si è occupato della questione, il quale ha suggerito, relativamente a disegni più generali di quello sopra descritto, una relazione approssimata definita dall'espressione:

$$deff = 1 + roh\,(\bar{m} - 1) \qquad (50)$$

in cui $\bar{m}$ indica il numero medio di US per UP e roh rappresenta il coefficiente di omogeneità, che costituisce un'estensione del coefficiente di correlazione intraclassi.

Tale relazione, sin dai primi anni settanta, ha assunto un ruolo di grande importanza in considerazione del fatto che si è rivelata uno strumento utilissimo soprattutto per la programmazione di nuove indagini campionarie; peraltro, è indispensabile il suo uso in altri campi di ricerca (cfr. nota 1). A tal proposito riteniamo utile sottolineare che:

— l'utilizzazione della (50) viene realizzata nei seguenti due modi: i) il primo, consiste nel determinare deff essendo noto roh; ii) il secondo, nel determinare roh essendo noto deff.

— la notevole utilità pratica della relazione in esame deriva dal fatto che gli indici deff e roh hanno il pregio di essere «portabili» (Kish, Groves e Krotki, 1976); hanno, cioè, la proprietà di essere usati per la stessa variabile in un contesto di indagine diverso da quello in cui sono stati calcolati. Addirittura, quando non è noto il valore appropriato di roh (oppure di deff) si può imputare il valore di roh (o di deff) di una variabile diversa da quella di interesse, purché appartenga alla stessa classe di variabili.

Nel corso di questi ultimi anni, tuttavia, si riscontra un rinnovato interesse verso questa importante relazione, concepita, a nostro avviso nel contesto di una filosofia del buon senso, dell'approssimazione e del compromesso fra rigore metodologico e praticità empirica.

L'obiettivo comune dei contributi più recenti è quello di una fervida ed appassionata ricerca di una relazione più generale della (50), che consenta di esprimere in modo meno approssimativo il legame tra deff e roh.

Su una linea di ricerca, in parte coincidente con quella appena delineata, anche nel nostro Reparto abbiamo avviato un programma di studi il cui obiettivo è la determinazione, relativamente al contesto campionario delle indagini sulle famiglie, di una relazione atta ad esprimere il legame tra deff e roh con un grado di approssimazione più spinto di quello offerto dalla relazione (50).

Un secondo obiettivo, che in un certo senso costituisce una naturale estensione di quello sopra formulato, riguarda la ricerca di relazioni che consentano la determinazione del coefficiente di omogeneità roh_s per le sottoclassi del campione, quando è noto il coefficiente roh di una data variabile per l'intero campione.

Anche questo secondo obiettivo è di grande importanza per il ruolo che roh_s riveste nel quadro dei problemi relativi alla predisposizione di future indagini e alla presentazione degli errori campionari.

Alcuni tentativi, in campo internazionale, sono stati fatti in tal senso, anche se i risultati, nelle stesse affermazioni degli Autori, devono considerarsi come largamente approssimativi. Ad esempio, in Kish ed altri (1976) si consiglia di moltiplicare roh per un coefficiente costante di 1,2 per ottenere roh_s.

Avendo, per grandi linee, delineato le tendenze che, a nostro parere, negli anni a venire potrebbero essere determinanti nell'orientare i futuri sviluppi dell'analisi statistica dell'effetto del disegno di campionamento, concludiamo la presente nota svolgendo alcune considerazioni sulle ricerche empiriche, sommariamente illustrate in questa nota.

Le riflessioni, concettuali e metodologiche, fatte nel corso di tali ricerche hanno consentito di convincerci su alcuni aspetti di fondo, concernenti i più importanti problemi che si incontrano in tema di programmazione e formazione di disegni campionari: scelta delle

variabili di stratificazione, determinazione di una soglia in base alla quale ripartire i comuni nei due domini territoriali, autorappresentativo e non autorappresentativo, calcolo delle numerosità campionarie di primo e di secondo stadio, definizione dei criteri per la ripartizione delle numerosità complessive di primo e di secondo stadio negli strati elementari.

In questa nota, per ovvie ragioni, limitiamo le considerazioni soltanto a due di tali problemi.

Anzitutto, quello relativo alla stratificazione. Alla luce dei valori dell'effetto stratificazione, illustrati precedentemente e di quelli ottenuti con altre ricerche qui non commentate, è nostro convincimento che è poco realistico pensare di poter individuare un insieme di variabili di stratificazione che consentano di ottenere guadagni nella precisione delle stime molto più elevati di quelli già descritti. Peraltro, questa conclusione sembra in accordo con i risultati ottenuti con riferimento ad indagini condotte in altri Paesi. Per questa ragione e per altre considerazioni di natura metodologica abbiamo preso la decisione di utilizzare la sola dimensione demografica dei comuni per stratificare gli stessi nell'ambito dei dominio non autorappresentativo. Un secondo aspetto di fondo, legato ai valori dell'effetto ponderazione, riguarda l'individuazione di una più efficace distribuzione delle famiglie campione negli strati elementari, in quanto come si è visto, l'elevata variabilità dei coefficienti di ponderazione produce aumenti delle varianze campionarie delle stime delle indagini. In considerazione di questo fatto abbiamo ritenuto opportuno nella formazione dei campioni imporre il vincolo dell'autoponderazione dei valori campionari, per l'ottenimento di stime centrate delle caratteristiche della popolazione oggetto di indagine.

NOTE

(1) Il notevole interesse manifestato per questo filone di ricerca è dovuto al fatto che l'effetto del disegno e le sue componenti, oltre a descrivere in modo più approfondito la natura delle variabili oggetto di studio, sono di grande importanza:

— nella valutazione critica dei piani di campionamento utilizzati per l'effettuazione di indagini effettive (Napolitano, Russo e Zannella, 1983; Russo, 1984; Verma, Scott e O'Muircheartaigh, 1980);

— nella predisposizione di future indagini (Kish, 1965; Kish, Groves e Krotki, 1976; Fabbris, 1980);

— nella costruzione di modelli per la presentazione degli errori di campionamento (Verma, 1982; Russo 1987);

— nell'analisi dei dati provenienti da campioni complessi (Kish e Frankel, 1974; Fuller, 1975; Fellegi, 1980; Holt, Scott e Ewings, 1980; Rao e Hidiroglou, 1981).

(2) I primi studi sono dovuti a Fabbris (1980).

(3) È opportuno precisare che l'autoponderazione è generalmente adottata nelle indagini condotte sulle famiglie.

(4) Tenendo presente il meccanismo probabilistico di selezione dell'unità, il peso k_{hij} è definito dall'espressione:

$$k_{hij} = \frac{N_h}{n_h} \quad \frac{M_{hi}}{m_{hi}}$$

(5) Il peso k_{ij} si ottiene immediatamente dall'espressione di cui alla nota (4), eliminando l'indice di strato, cioè:

$$k_{hj} = \frac{N}{n} \quad \frac{M_i}{m_i}$$

(6) $$k_{hj} = \frac{M_h}{m_h}$$

(7) I valori degli effetti indicati, per gruppi di variabili, sono calcolati come rapporto tra errori campionari, anziché tra varianze.

RIFERIMENTI BIBLIOGRAFICI

COCCIA G., (1986), *Un metodo per la stima dell'effetto complessivo di campionamento nei campioni complessi*, Atti della XXXIII Riunione della SIS, Bari.

COCCIA G., D'ANGIOLINI, G., FALORSI, P. e RUSSO A. (1987), *Una metodologia per la valutazione degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo del disegno di campionamento nell'indagine sulle forze di lavoro*, Atti del Seminario su: «Forze di lavoro: Disegno dell'indagine ed analisi, strutturali», Dipartimento di Scienze Statistiche - Università di Padova, Bressanone.

FABBRIS L., (1980), *Problemi statistici per il sovracampionamento su base nazionale delle forze di lavoro*. Convegno internazionale su «L'informazione statistica su scuole e mercato del lavoro e sulle politiche per l'occupazione giovanile», Fondazione G. Cini, Venezia.

FALORSI S. e FALORSI P. (1989), *A method for the estimation of design effect in a two stage stratified sample survey with PSU selection without replacement and unequal probabilities, in corso di stampa negli Atti del Convegno ISI*, Parigi.

FELLEGI I.P., (1980), *Approximate Tests of Independence and Goodness of Fit Base on Stratified Multistage Sample*, Journal of the American Statistical Association, Vol. 75.

FULLER W.A., (1970), *Sampling with random stratum boundaries*, Journal of the Royal Statistical Society, B. 32.

FULLER W.A., (1975), *Regression Analysis for Sample Surveys*, Sankhya, Serie C. Vol. 37.

HOLT D., SCOTT, A.J. e EWINGS P.O., (1980), *Chi-Scuared Test with Survey Data*, Journal of the Royal Statistical Society, Sec. A, 143.

KISH L. (1965), *Survey Sampling*, Wiley, New York.

KISH L. FRANKEL M.R. (1974), *Inference from Complex Samples*, Journal of the Royal Statistical Society, B, 36.

KISH L. GROVES R.M. e KROTKI K.P. (1976), *Sampling Errors for Fertility Surveys*, WFS Occasionale Papers, n. 17.

NAPOLITANO P. RUSSO A. e ZANNELLA F. (1983), *Calcolo, presentazione ed analisi degli errori di campionamento nella Indagine Istat sulle condizioni di salute della popolazione e sul ricorso ai servizi sanitari*, Atti del Convegno della SIS, Trieste.

RUSSO A. (1984), *Calcolo ed analisi degli errori di campionamento nell'Indagine Istat sulle vacanze e gli sports degli italiani, anno 1982*, Atti della XXXII Riunione della SIS, Sorrento.

RUSSO A. (1985), *Su un metodo di stima degli effetti stratificazione, clustering e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con stratificazione delle unità di primo stadio*, Quaderni di Discussione, n. 5, Istat, Roma.

RUSSO A. (1986a), *Su un metodo di stima dell'effetto ponderazione nei campioni a due stadi con stratificazione delle unità primarie,*, Quaderni di Discussione, n. 1 Istat, Roma.

RUSSO A. (1986b), *Un metodo di stima dell'effetto della stratificazione nei campioni complessi*, Atti della XXXIII Riunione della SIS, Bari.

RUSSO A. (1986c), *Una metodologia per la stima degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con selezione delle unità primarie con reimmissione e probabilità variabili*, Quaderni di Discussione n. 2 Istat, Roma.

RUSSO A. (1987), *Sulla presentazione degli errori di campionamento mediante modelli: il metodo dei modelli regressivi*, Quaderni di Discussione, n. 4 Istat, Roma.

RUSSO A. (1988), *Un metodo per la stima dell'effetto della post-stratificazione nei campioni a due stadi*, Atti della XXXIV Riunione delle SIS, Siena.

VERMA V., SCOTT C. e O'MUIRCHEARTAIGH (1980), *Sample Design and Sampling Errors for the World Fertility Survey*, Journal of the Royal Statistical Society, A, Part. 4.

VERMA V. (1982), *The Estimation and Presentation of Sampling Errors*, Technical Bulletins, W.F.S, New York.

IL SISTEMA DI CONTROLLO DELLE INDAGINI CAMPIONARIE DELL'ISTAT: LINEE DI RICERCA E PRINCIPALI CONTRIBUTI DEL PROGETTO «QUALITÀ DI DATI»

di *Mauro Masselli, Marina Signore* (*)

1. L'INDAGINE STATISTICA E LA QUALITÀ DEI DATI

L'informazione prodotta da una indagine statistica è costituita, oltre che dai risultati di elaborazioni sui dati elementari (tavole, indicatori ecc.), anche dagli stessi microdati (sotto forma di archivi) e da tutte quelle notizie necessarie ad una loro corretta utilizzazione (definizioni, classificazioni, questionario, notizie sulla rilevazione sul campo ed indicatori di qualità).

La valutazione della qualità dei risultati dell'indagine non può, quindi, essere limitata ai soli macrodati, ma deve essere estesa a tutti e tre i succitati livelli di informazione, microdati, macrodati e metadati; di conseguenza, è opportuno tentare di definire il concetto di qualità in modo da comprendere i vari aspetti considerati e, per evitare ambiguità, fare riferimento alla «qualità dell'informazione prodotta» più che alla «qualità dei dati».

In termini generali, possiamo definire come «qualità di un prodotto», l'adeguatezza del prodotto all'uso, ovvero la capacità di un prodotto a soddisfare le proprietà garantite dal produttore (D.G. Montgomery 1985; O.Arkhipoff, 1986). Le proprietà, garantite implicitamente od esplicitamente dal produttore, possono essere suddivise in due insiemi: i) di progettazione e ii) di tolleranza.

Le prime riguardano le caratteristiche proprie del prodotto-tipo, mentre le seconde si riferiscono all'affidabilità di tali caratteristiche nei differenti «pezzi» effettivamente prodotti.

Adattando questa definizione al caso di una indagine statistica, possiamo riferire le garanzie di progettazione alla capacità dell'indagine di soddisfare gli obiettivi conoscitivi per cui è stata condotta; in particolare (a) se il tempo intercorso tra la progettazione e la disponibilità dei risultati non ne ha intaccato la validità e l'utilità (tempestività); (b) se gli obiettivi sono stati ben individuati (rilevanza teorica); (c) se i microdati ed i macrodati derivanti dalle elabo-

razioni sono congrui con tali obiettivi (rilevanza effettiva); (d) se le informazioni accessorie, i metadati, sono sufficienti per interpretare correttamente i risultati (trasparenza).

Tali «garanzie» non caratterizzano completamente la qualità dell'informazione; come nel caso dei manufatti, si deve considerare l'aspetto di «tolleranza» del prodotto, ovvero in termini statistici, di «precisione» delle stime ottenute.

La «precisione» di una stima è definita come funzione inversa dell'errore statistico, (differenza tra valore «osservato» e valore «vero»); l'errore può derivare o dalla tecnica campionaria, ossia dall'aver raccolto dati solo in una parte dell'universo considerato (errore campionario), oppure dalla discrepanza tra le condizioni ideali in cui si sarebbe dovuta svolgere l'indagine e quelle effettivamente realizzate (errore non campionario). Le garanzie di tolleranza, quindi, saranno definite in termini di (e) precisione campionaria e (f) precisione non campionaria.

In sintesi, possiamo caratterizzare la qualità dell'informazione, derivante da una indagine statistica, mediante:

— le garanzie di progettazione:

(a) tempestività
(b) rilevanza teorica
(c) rilevanza effettiva
(d) trasparenza

— le garanzie di tolleranza:

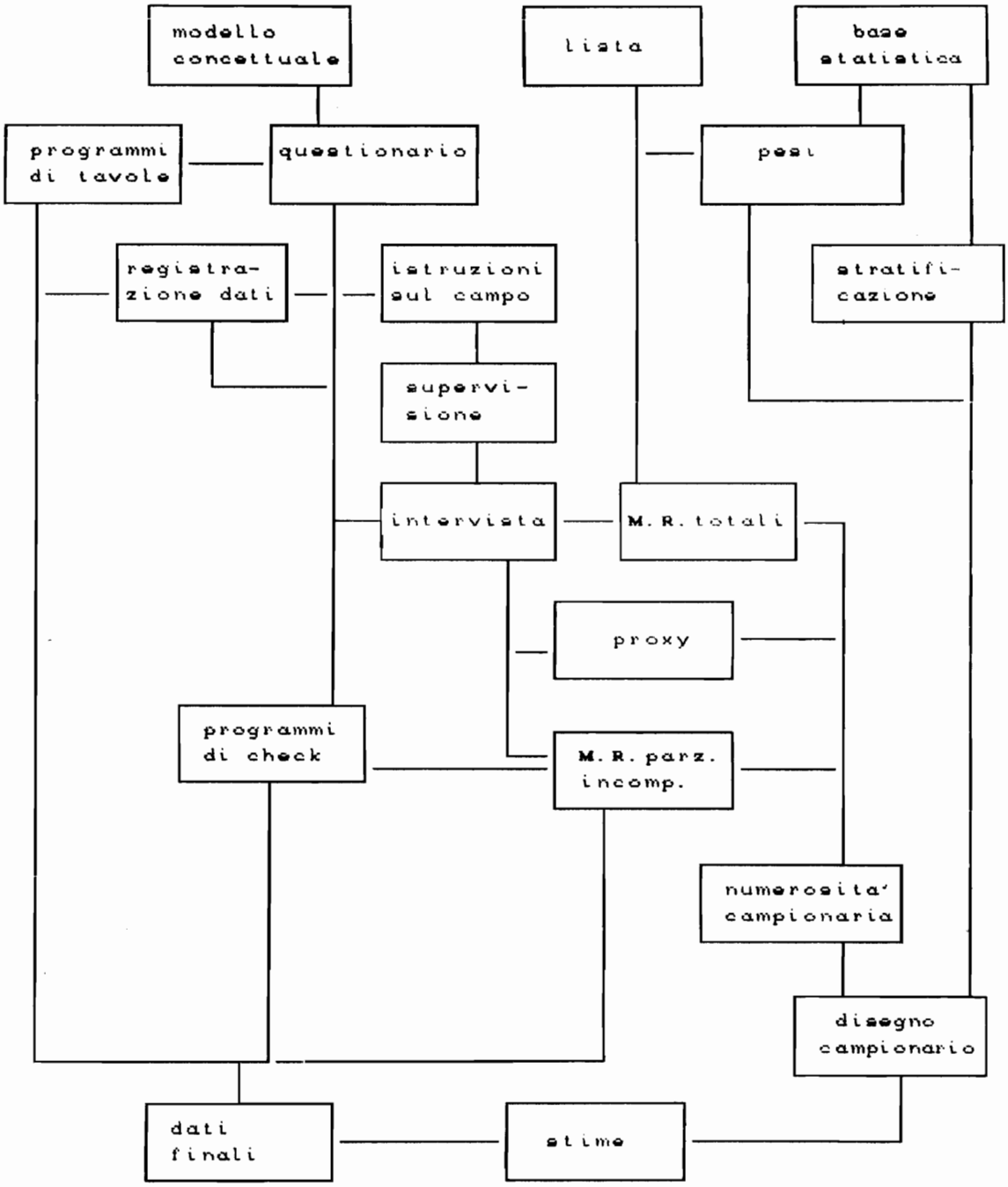
(e) precisione campionaria
(f) precisione non-campionaria

Nella definizione di qualità sopra riportata, sono presenti elementi relativi e soggettivi; le «garanzie» non vengono determinate da standard teorici, validi sempre e comunque, bensì sono fissate in funzione dei costi/benefici derivanti dall'informazione prodotta.

2. LA SCOMPOSIZIONE DELL'ERRORE TOTALE

Le «garanzie di tolleranza» riguardano l'errore globale della rilevazione statistica, che viene prodotto dall'interazione degli errori generati nelle diverse operazioni, necessarie alla conduzione dell'in-

Figura 1


modello concettuale
lista
base statistica
programmi di tavole
questionario
pesi
registrazione dati
istruzioni sul campo
stratificazione
supervisione
intervista
M. R. totali
proxy
programmi di check
M. R. parz. incomp.
numerosita' campionaria
disegno campionario
dati finali
stime

dagine; in figura 1 sono state schematizzate le principali relazioni che legano le fasi del processo e quindi i legami tra i relativi errori.

L'errore globale può manifestarsi sia come distorsione rispetto al «valore vero», sia come aumento della variabilità dello stimatore e può essere misurato dalla media quadratica della distanza tra il generico stimatore y ed il valore «vero» Y; l' MSE (y), può quindi essere scomposto in una parte variabile V (y) ed in una distorsione B (y):

$$MSE\ (y) = E\ (y - y)^2 = V\ (y) + B^2\ (y) \qquad (1)$$

Se ci si riferisce a differenti tipologie di errore la (1), può essere riscritta come:

$$MSE\ (y) = \Sigma_i\ V\ (y_i) + \left[\Sigma_i\ B\ (y_i)\right]^2 + 2 \sum_{i>j} cov\ (y_i\ y_j) \qquad (2)$$

La (2) è del tutto generale e rappresenta l'errore totale come somma di distorsioni, varianze e covarianze derivanti dalle diverse fonti, cui è riferito l'indice; in tale rappresentazione, quello campionario è considerato come uno dei diversi tipi d'errore.

Le due tipologie di errore, tuttavia, differiscono sostanzialmente; l'errore campionario, dovuto alla selezione di parte delle unità della popolazione in esame, dipende dalla variabilità del fenomeno in studio, dalla strategia di campionamento e dagli stimatori adottati, mentre quello non campionario è funzione della soggettività del ricercatore, degli aspetti organizzativi della rilevazione, del comportamento di una pluralità di soggetti e del contesto socio-culturale in cui si colloca l'indagine.

Tale distinzione si riflette anche sulla «controllabilità», e persino sulla «misurabilità», dell'errore; infatti, mentre il primo dipende dalla scelta del disegno di campionamento ed è calcolabile mediante acquisizioni teoriche ben collaudate, il secondo è solo parzialmente influenzabile dalle scelte pratiche ed organizzative effettuate dal produttore e, nella maggior parte dei casi, è difficile ottenerne una stima sintetica.

Cosicché mentre il controllo dell'errore campionario avviene mediante la programmazione del disegno di campionamento, quello non campionario è specificatamente l'oggetto del «controllo statistico di qualità».

3. IL SISTEMA DI CONTROLLO

L'indagine può essere assimilata ad un processo di produzione manifatturiero; la materia prima, l'informazione indistinta in possesso delle unità di rilevazione, viene «trattata» in numerose fasi (rilevazione, revisione, elaborazione) ed operazioni interrelate, fino al prodotto finale.

Garantire la qualità dell'informazione significa, allora, garantire gli standard del processo di produzione mediante il «controllo» delle differenti operazioni. Date le interrelazioni e le connessioni tra queste ultime, l'efficacia dei controlli è in relazione all'integrazione ed all'organicità con cui essi vengono programmati e realizzati; i controlli, devono essere progettati come «sistema».

Il sistema di controllo ha per oggetto l'errore non campionario e per obiettivi:

i) la prevenzione dell'errore
ii) la correzione dell'errore
iii) la stima dell'errore totale
iv) il monitoraggio delle fasi del processo di formazione del dato.

Elementi costitutivi del sistema sono la definizione delle fasi del processo e dei livelli di controllo ritenuti necessari, la programmazione delle fonti e dell'organizzazione dell'informazione sugli errori, i metodi di prevenzione, analisi e correzione.

Per quanto riguarda le fasi, possiamo suddividere l'indagine nei seguenti blocchi di operazioni omogenee, sotto il profilo temporale ed organizzativo, nel flusso del processo di produzione:

— progettazione dell'indagine
— rilevazione sul campo
— codifica
— registrazione su supporto informatico
— revisione e correzione

— elaborazione dei risultati
— validazione e diffusione dei risultati

In ciascuna fase possono essere generati errori che si elidono o si combinano con quelli derivanti dalle fasi precedenti; l'esistenza di queste interrelazioni, rende l'attribuzione di un particolare errore ad una sola fase od operazione, un artificio logico, necessario per la classificazione e per la modellizzazione dell'errore, ma non rispondente alla realtà.

Nella fase di programmazione dell'indagine, possono venir generati errori di rilevanza teorica nella definizione degli obiettivi, delle definizioni e delle classificazioni; errori di rilevanza effettiva nella predisposizione del piano di diffusione dei risultati; errori di misura nella programmazione delle differenti operazioni (necessarie per la raccolta sul campo, per la revisione e correzione del materiale «grezzo» e per l'elaborazione dei risultati) e nella loro integrazione in un unico complesso organizzativo.

Nella fase di rilevazione sul campo, i diversi «operatori» (supervisori, rilevatori, rispondenti, l'assistenza centralizzata alla rete) e le loro interrelazioni, possono causare errori di misura; i medesimi errori sono attribuibili alle operazioni di codifica e registrazione su supporto informatico.

Alla fase di revisione e correzione è demandato il compito di determinare gli errori e di correggerli; tuttavia, è possibile che le procedure non riescano a identificarli o ne generino di propri.

Nell'elaborazione dei macrodati, possono essere indotti errori dovuti al calcolo o imputabili ai coefficienti di espansione; infine la validazione può risultare insufficiente ad assicurare la coerenza dei risultati.

I «livelli» vengono definiti come le operazioni, o gli operatori, che possono generare errori (ad es. il rilevatore, il comune, la registrazione etc.) e che vengono effettivamente sottoposti a controllo; essi costituiscono un sottoinsieme delle possibili fonti di errore, prescelto sulla base di considerazioni di ordine organizzativo ed economico.

In termini operativi, la fase rappresenta il punto nel flusso logico-temporale della produzione in cui è possibile o conveniente effettuare il controllo (si potrebbe dire il «quando»), mentre il livello è l'opera-

zione su cui il controllo viene esercitato (il «dove»). Il medesimo livello può essere controllato in fasi differenti e in ciascuna fase possono essere controllati più livelli.

Per le indagini campionarie sulla popolazione, condotte dall'Istat, le fasi, i livelli ed il relativo tipo di errore sono riportate nel Prospetto 1.

Prospetto 1

fasi/operazione	livelli	tipo di errore
progettazione	— modello concettuale	— rilevanza
questionario	— struttura — classificazioni — definizioni — lunghezza — vocabolario	— rilevanza — misura
selezione PSU	— base statistica	— calcolo probabilità di inclusione
selezione SSU	— base statistica — lista	— calcolo probabilità di inclusione — copertura
formazione elenchi ed assegnazione	— supervisori — rilevatori	— identificazione delle unità
rilevazione sul campo	— supervisori — rilevatori — rispondenti	— M.R. totali — M.R. parziali — incongruenze — proxy — memoria
registrazione	— operatori	— valori erronei o mancanti
revisione e correzione	— revisori — pianti di compatibilità	— misura
stima	— base statistica	— calcolo fattori di espansione
elaborazione	— programmi	— calcolo

Definite le fasi ed i livelli, il passo successivo consiste nell'individuare le fonti informative, necessarie per la determinazione dell'errore, e del relativo sistema di riconoscimento e di collegamento tra informazioni riguardanti la medesima unità. Le fonti di informazione per il controllo dell'indagine, sono molteplici e derivano sia dagli strumenti di rilevazione ed elaborazione dei dati (il questionario, i risultati del piano di compatibilità e correzione, i risultati del controllo della registrazione, le indagini di controllo), sia dalla documentazione amministrativa e di supporto (piano di campionamento, gli elen-

chi dei rilevatori e le relative assegnazioni, le informazioni sulle mancate risposte totali e sulle unità sostituite), sia, infine, dalla comparazione con fonti esterne.

Tale ampia base informativa deve essere organizzata nel «Sistema Informativo Statistico» dell'indagine, per cui diviene rilevante l'affidabilità e la completezza del sistema di codici identificativi delle unità che deve assicurare il collegamento tra archivi: esso deve essere progettato in funzione degli obiettivi di controllo prefissati.

Gli indicatori che è possibile costruire dalle informazioni contenute dal data-base, costituiscono l'«archivio di qualità»: essi verranno definiti indicatori di un determinato «livello», se disponibili per tutte le unità appartenenti allo stesso. Potendo, inoltre, ordinare gerarchicamente almeno una parte dei livelli, mediante la relazione di inclusione delle relative unità (ad esempio: individuo $\subset$ famiglie $\subset$ rilevatore $\subset$ comune $\subset$ aggregazioni di comuni), l'analisi di livelli gerarchicamente superiori può giovarsi, oltre che di proprie fonti informative, anche di aggregazione di indicatori derivanti dal quelli inferiori.

L'archivio di qualità costituisce, dunque, la base informativa per rispondere ad esigenze diverse:

— lo studio del «profilo» dell'errore dell'intera rilevazione,
— l'analisi della procedura per singola fase,
— la programmazione di indagini similari o per la stessa indagine in tempi differenti.

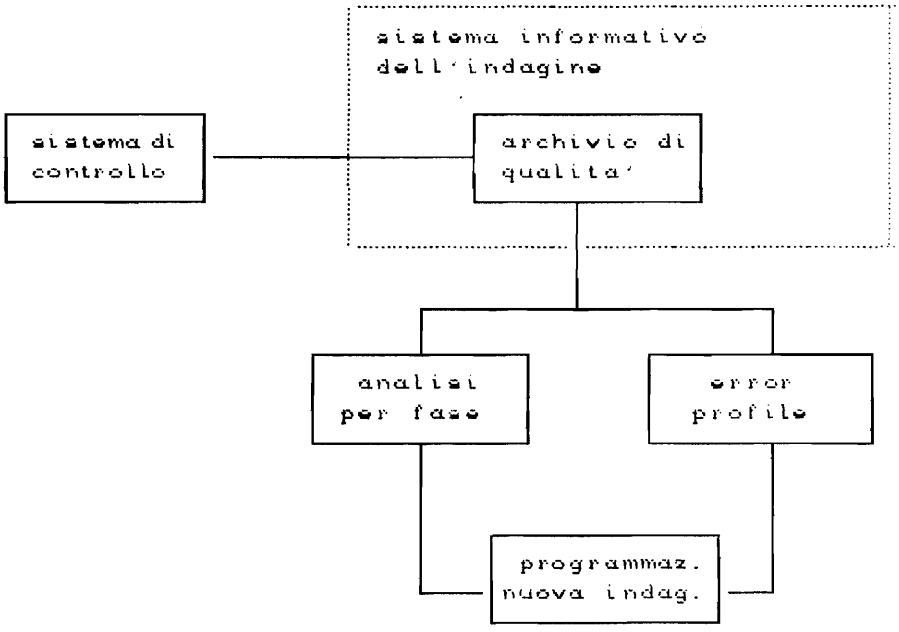
Tale ultima utilizzazione è particolarmente efficiente in caso di indagini ripetute o che insistono sulla medesima rete di rilevazione; infatti, mediante l'analisi dell'archivio, si possono determinare le situazioni anomale, su cui concentrare gli sforzi organizzativi, invece di disperdere le risorse con interventi «a pioggia».

Nella Figura 2 vengono schematizzate le relazioni tra «sistema di controllo», «sistema informativo statistico», «archivio di qualità» e le possibili utilizzazioni di tale archivio.

Il controllo di qualità viene realizzato mediante analisi differenti che possono essere classificate nelle seguenti categorie, in funzione della provenienza e del livello di aggregazione dell'informazione:

— analisi dei dati desumibili dal processo di rilevazione e di elaborazione;

Figura 2


sistema informativo
dell'indagine
sistema di
controllo
archivio di
qualita'
analisi
per fase
error
profile
programmaz.
nuova indag.

— confronto con risultati aggregati, ottenuti da altre indagini o dalla stessa in tempi diversi;

— confronto con micro-dati, provenienti da altre indagini o dalla stessa in tempi diversi;

— risultati di indagini di controllo.

In generale, mentre i primi tre metodi richiedono una revisione delle attuali procedure e competenze a livello centrale, l'ultimo implica una ripetizione della rilevazione su sub-campioni e, quindi, un maggiore aggravio, sia in termini economici che organizzativi.

Una ulteriore distinzione può essere stabilita rispetto al momento in cui avviene il controllo ed ai fini per cui si effettua: controllo preventivo, nel caso che preceda la rilevazione sul campo, con l'obiettivo di verificare e migliorare la programmazione dell'indagine; controllo in corso d'opera, con l'obiettivo di correggere l'errore riscontrato; controllo successivo, finalizzato alla stima ed all'analisi dell'errore.

Mediante l'archivio di qualità, per le indagini ripetute e per quelle in cui può essere ipotizzata la «portabilità» degli indicatori, i controlli «successivi» della singola rilevazione assumono il ruolo di controlli «preventivi» nei confronti delle rilevazioni successive.

4. LA PREVENZIONE DELL'ERRORE

Le tecniche per prevenire l'errore sono fondamentali per ridurre l'errore totale; infatti, come verrà meglio specificato nel paragrafo 5, i metodi di correzione risolvono solo parzialmente, ed in modo non del tutto soddisfacente, il problema.

In sede di programmazione, quindi, è necessario eliminare l'errore di «rilevanza», mettere a punto lo strumento tecnico di rilevazione, il questionario, e l'intero complesso dell'indagine. Il controllo del modello di rilevazione viene effettuato con varie tecniche: il giudizio degli esperti, il pre-test sul campo ed il test di alternative. Infine per il controllo dell'intera indagine si ricorre all'indagine pilota.

4.1 La progettazione concettuale

I contenuti informativi di una indagine statistica possono essere individuati e rappresentati in maniera formale, indipendentemen-

te dalle problematiche specifiche. Il modello «Entità Relazioni» permette di definire e documentare le relazioni tra entità, gli attributi delle stesse e le strutture gerarchiche tra entità. L'utilizzo del modello costringe a definire in maniera precisa i concetti coinvolti; tali definizioni costituiscono i metadati dell'indagine. La documentazione sulle definizioni e la rappresentazione formale degli schemi concettuali, costituiscono il patrimonio informativo dell'indagine e permettono il controllo della rilevanza, teorica ed effettiva, dell'informazione prodotta.

Il modello concettuale può essere utilizzato per la redazione del questionario; è infatti possibile estrarre da esso un «albero di aree omogenee» di informazione, che sarà tradotto nella struttura e nei quesiti del questionario.

Poiché, attraverso il modello concettuale è possibile rappresentare le relazioni tra le diverse entità ed i loro attributi, gli schemi prodotti possono essere utilizzati anche per definire l'insieme delle regole di compatibilità, mediante le quali vengono determinate le incongruenze logiche dei dati raccolti.

4.2 Il test di alternative

Il test consiste nel sottoporre a verifica, su campioni bilanciati, due, o più, versioni del questionario, che differiscono per un aspetto (ad es. sequenza delle domande, formulazioni di quesiti, periodi di riferimento temporali, la sequenza delle domande etc.); tale tecnica guida nella scelta tra due alternative che appaiono equivalenti, in mancanza di altre informazioni.

4.3 Il pre-test del questionario

Una volta stabiliti i contenuti informativi, e redatta una versione provvisoria del questionario, il medesimo è sottoposto a verifica sul campo, mediante somministrazione ad un campione di unità. Il campione deve rispecchiare la massima variabilità delle condizioni della rilevazione e delle caratteristiche strutturali delle unità; per questo motivo è conveniente utilizzare campioni ragionati e non probabilistici. I rilevatori utilizzati devono essere selezionati tra il personale particolarmente esperto e formati con particolare cura; ad essi

vanno richieste, con apposite riunioni, tutte quelle informazioni difficilmente quantificabili in modelli aggiuntivi e riguardanti:

— la completezza e la correttezza del questionario rispetto agli obiettivi;

— le difficoltà riscontrate dagli intervistati ed il loro atteggiamento di fronte all'indagine;

— la semplicità di gestione da parte dell'intervistatore dello strumento «questionario».

4.4 L'indagine pilota

L'indagine pilota si differenzia dal pre-test del questionario in quanto ha l'obiettivo di verificare tutti gli aspetti delle rilevazione ed è condotta mediante un campione probabilistico.

L'indagine pilota costituisce una «versione ridotta» dell'indagine principale di cui verifica tutte le procedure; d'altro canto essa dovrebbe anche comportare operazioni più accurate e controllate in modo da identificare gli errori. Si può quindi affermare che l'indagine pilota è meno «estesa» ma più «approfondita» rispetto all'indagine madre; per suo mezzo si raccolgono non solo le caratteristiche oggetto di studio (allo scopo di stimare la variabilità dei fenomeni e quindi determinare, in mancanca di altre fonti, la numerosità campionaria) ma anche le informazioni concernenti l'organizzazione dell'indagine. A tale scopo è conveniente associare alla pilota un corpo selezionato di «supervisori» e prevedere modelli ad hoc e relazioni per il controllo delle procedure ai vari livelli e fasi.

5. LA CORREZIONE DELL'ERRORE

Gli errori che è possibile correggere, nella pratica di indagini di media-grande dimensione, sono sostanzialmente quelli derivanti dalle mancate risposte totali e parziali, e quelli di coerenza della singola variabile o tra variabili logicamente collegate; sono questi infatti gli errori identificabili in tempi utili, per non mutare le condizioni generali di svolgimento dell'indagine, e con costi economici ed organizzativi contenuti.

Da quanto detto, discende che l'errore che può essere corretto è solo una parte dell'errore totale.

Assimilando le coerenze «fallite» a mancate risposte parziali, possiamo definire due sole tipologie di errore, le mancate risposte totali e quelle parziali; a tali tipologie corrispondono differenti possibilità di correzione.

5.1 Le mancate risposte totali

Le mancate risposte totali producono una distorsione delle stime, se il meccanismo che le genera è, come avviene generalmente nella realtà, non casuale. In questo caso la distorsione è funzione della quota di non rispondenti nelle diverse sub-popolazioni, e della differenza tra i parametri dei rispondenti e dei non rispondenti. Ad esempio, nel caso di una media la distorsione dello stimatore è data dalla:

$$B(\bar{y}_R) = \Sigma_i B(\bar{y}_{iR}) = \Sigma_i E(\bar{y}_{iR} - \bar{Y}_i) = \Sigma_i W_{iR}(\bar{Y}_{iR} - \bar{Y}_{iNR}) \qquad (3)$$

dove:

i = indice di sub-popolazione
$\bar{y}_R$ = media campionaria dei rispondenti
$\bar{y}_{iR}$ = media campionaria dei rispondenti nella i-esima sub-popolazione
$\bar{Y}_{iR}$ = media dei rispondenti nella i-esima sub-popolazione
$\bar{Y}_{iNR}$ = media dei non rispondenti nella i-esima sub-popolazione
$\bar{Y}_i$ = media nella i-esima sub-popolazione
W_{iR} = quota dei rispondenti nella i-esima sub-popolazione

In questo caso, la correzione avviene a livello di stime, modificando, con appositi pesi, le probabilità finali di inclusione delle unità campionarie:

$$\bar{y} = \Sigma_{ij} P_{ij} y_{ij} \rightarrow \Sigma_{ij} \phi_{ij} y_{ij}$$

La riduzione della distorsione, operata da tale procedimento, è funzione dell'omogeneità delle sub-popolazioni individuate a posteriori.

5.2 Le mancate risposte parziali

La correzione delle mancate risposte parziali, al contrario di quelle totali, avviene a livello dei micro-dati rilevati, mediante due possibili tecniche: il ritorno presso l'unità rispondente e le operazioni di revisione effettuate sul materiale raccolto.

L'attuale organizzazione e la dimensione delle indagini, non permettono, a tutt'oggi, il ricorso alla prima tecnica in maniera diffusa e totale; cosicché la revisione rimane una fase insostituibile nelle rilevazioni condotte dall'Istituto.

L'identificazione e la correzione degli errori presenti nei dati «grezzi», può essere condotta o mediante «esperti» di settore, o avvalendosi di programmi informatici, o mediante un mix delle due tecniche. Il vantaggio nell'utilizzare procedure informatiche risiede nella maggiore tempestività, nella controllabilità delle operazioni effettuate, nella omogeneità di trattamento dell'errore.

5.3 I programmi di compatibilità e correzione

I programmi di compatibilità e di correzione hanno la duplice funzione di determinare le incongruenze e correggerle; tali funzioni sono logicamente distinte, anche se la maggior parte degli algoritmi le effettua simultaneamente.

L'identificazione dell'errore viene assicurata da un insieme di regole, basate:

— sui valori non ammissibili per le singole variabili;
— sulle relazioni logiche tra variabili;
— sulla sequenza del questionario.

La costruzione dell'insieme di regole è un'operazione complessa; le regole esplicitate, infatti, possono essere ridondanti, o possono, implicitamente, definire nuove regole contraddittorie. Le prime hanno l'effetto di gravare sui tempi di esecuzione, mentre le seconde inficiano l'intera operazione di correzione. Per evitare tali inconvenienti, si può fare ricorso ai risultati della progettazione concettuale o ad un algoritmo che garantisce la costruzione di un insieme minimo, non ridondante e non contraddittorio, a partire dalle regole esplicitate (Fellegi - Holt, 1976).

I criteri di correzione sono vari e non è agevole darne una classificazione esaustiva e precisa; per di più vengono spesso usati in combinazione tra loro. Si possono, comunque, distinguere i principali metodi in:

— deterministici;
— da donatore, ulteriormente distinguibili in «hot-deck» e «cold-deck»;
— da regressione, deterministica o stocastica.

Gli algoritmi deterministici rispondono ad una logica «IF-THEN» e sono assimilabili ad un «albero» di decisione. Essi implicano un ordinamento gerarchico tra le variabili e determinano e correggono l'errore della i-esima, sulla base del valore assunto dalle k variabili precedenti. La validità delle correzioni di tale procedura, dipende dalla gerarchia prescelta e dalla probabilità di errore delle singole variabili.

Nelle procedure «da donatore», le variabili da correggere vengono sostituite con i valori assunti in una unità «donatrice», in cui non è stato riscontrato alcun errore. Il metodo «cold-deck» si differenzia da quello «hot-deck», in quanto, nel primo, le unità vengono preventivamente suddivise in due insiemi (senza errori e con almeno un errore); il secondo, invece, aggiorna continuamente un sottoinsieme, di dimensioni date, di unità «pulite», da cui preleva il donatore, sfruttando in questo modo le eventuali correlazioni esistenti tra unità «vicine» secondo un ordinamento indotto nel file. La popolazione donatrice, per una determinata unità, può essere costituita da tutte le unità senza errori, ovvero da gruppi selezionati mediante alcune caratteristiche; tali variabili (variabili di collegamento) non devono essere sottoposte a correzione e devono risultare altamente correlate tra loro.

Il criterio «hot-deck» è applicabile sia a caratteristiche qualitative che quantitative; per queste ultime, però, è necessario specificare una funzione di distanza per le variabili di collegamento e «perturbare» il dato del donatore, per evitare un eccessivo «appiattimento» della distribuzione.

Dati m rispondenti su n casi, la scelta del donatore può essere effettuata:

i) mediante selezione casuale senza reimmissione per $m > 1/2\ n$ (SR nel Prospetto 2);

ii) mediante selezione casuale con reimmissione per m < 1/2 n (CR nel Prospetto 2);

iii) mediante selezione casuale del più vicino rispondente (SEQ nel Prospetto 2);

iv) mediante selezione da file ordinato (ORD nel Prospetto 2);

v) mediante stratificazione delle unità e quindi operando con uno dei criteri (i) - (iv).

I quattro criteri di cui sopra, danno luogo a stimatori con varianze diverse; essi possono essere confrontati (Prospetto 2) con quello calcolato solo utilizzando le unità non errate, assimilate alle unità rispondenti.

Prospetto 2

stimatore	E (θ)	Var (θ)
$\bar{y}_R$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{n-m}{m}\right]$
$\bar{y}^{sr}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{2(n-m)}{n}\right]$
$\bar{y}^{cr}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \left(\frac{n-m}{m}\right)\left(\frac{n+m-1}{n}\right)\right]$
$\bar{y}^{seq}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{2(n-m)}{n}\right]$
$\bar{y}^{ord}$	$\bar{Y}_R$	$\frac{V}{n}\left[1 + \frac{2(n-m)}{m} + 2\left(\frac{\varrho}{1-\varrho} - \frac{n-m}{n} \cdot \frac{2\varrho}{1-\varrho}\right)\right]$

dove: $V = \text{Var}(y)$
$\varrho = \text{corr}(y_i, y_j)$ $i, j = 1, 2, \ldots n$

Il criterio di correzione mediante regressione è applicabile solo alle caratteristiche quantitative; esso consiste nello stimare il valore da sostituire, mediante l'usuale modello di regressione lineare:

$$\hat{y}_i = b_{r,o} + \Sigma_j\, b_{r,j}\, z_{i,j} + \hat{e}_i \qquad (4)$$

$\hat{y}_i$ = valore correttivo per l'unità i-esima

$b_{r,o}$, $b_{r,j}$ = coefficienti di regressione dei rispondenti nel j-esimo gruppo

$z_{i,j}$ = variabili ausiliari

$\hat{e}_i$ = residuo stocastico da aggiungere, con $E(\hat{e}_i) = 0$

Se il valore di $\hat{e}_i$ viene posto a zero, il modello di generazione è deterministico; stocastico nel caso contrario.

Operando sulle z_{ij} e sulle e_i, questo metodo equivale ad altre tecniche: i) se le variabili ausiliari ed il residuo sono poste a zero, esso coincide con la media dei rispondenti; ii) con le variabili ausiliarie uguali a zero ed i residui posti uguali alle differenze tra i valori dei rispondenti e la loro media, il metodo equivale alla selezione casuale di un rispondente; iii) se si effettua una stratificazione dei rispondenti considerando dummy le z_{ij}, si otterranno le medesime equivalenze di cui sopra a livello di singola classe. Il metodo di regressione stocastica, infine, è equivalente al criterio «hot-deck» sequenziale, da file ordinato e stratificato con selezione casuale.

6. LA STIMA DELL'ERRORE

L'errore non campionario rappresenta la componente più rilevante dell'errore totale di uno stimatore, la sua determinazione si rivela, quindi, indispensabile per conoscere la reale precisione delle stime fornite da un'indagine. D'altro canto la conoscenza dell'entità degli errori di misura, che si sono accumulati nelle varie fasi del processo di produzione del dato, è il presupposto necessario per il miglioramento della qualità dei dati rilevati.

La conoscenza della struttura dell'errore può essere finalizzata a più obiettivi:

— il monitoraggio del processo per singola fase;
— la stima dell'errore totale;
— la costruzione dell'error profile.

A tale scopo è possibile utilizzare indagini di controllo con specifici obiettivi, oppure ricorrere alle quantificazioni ottenute dalle procedure standard dell'indagine. Appartengono ad esempio a questa seconda categoria le informazioni sulle mancate risposte totali e parziali, i risultati dei piani di compatibilità, le informazioni riguardanti l'intervista ed i rispondenti, quelle relative al campione di controllo della registrazione, il carico di lavoro per intervistatore etc. Gli indicatori che è possibile calcolare confluiscono nell'archivio di qualità; da questo si può estrarre una matrice-dati, relativa a ciascuna fase o a ciascun livello di controllo, che può essere analizzata con le consuete metodologie univariate o multivariate.

6.1 La stima dell'errore mediante modelli

La quantificazione degli effetti degli errori di misura è conseguente alla formalizzazione del problema mediante un adeguato modello matematico. Tra i vari modelli presenti nella letteratura, il più noto ed utilizzato è quello introdotto da Hansen, Hurwitz e Bershad (1961), poi ripreso e generalizzato da vari autori tra i quali Fellegi (1964) e Koch (1973).

Il modello matematico di Hansen, Hurwitz e Bershad si basa sulle seguenti ipotesi:

— l'indagine è ripetibile sotto le stesse condizioni essenziali;
— le replicazioni del processo di misurazione sono indipendenti tra loro.

Consideriamo un campione di n unità estratte in blocco da una popolazione di N. Sia y_{ijt} il valore osservato per la i-esima unità (i = 1,2,...,n') dal j-esimo intervistatore (j = 1,2,...,k e n = n' k) nella t-esima replicazione e sia μ_i il valore «vero» relativo all'unità i. Indichiamo con m_i il valore atteso di y_{ijt} al variare della replicazione del processo di misurazione sull'unità i, con M la media nella popolazione di tali valori attesi e con μ quella dei valori «veri»; in simboli:

$$m_i = E (y_{ijt} \mid i) \tag{1}$$

$$M = \frac{1}{N} \sum_{i=1}^{N} m_i \tag{2}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mu_i \tag{3}$$

Si definiscono, quindi, la deviazione di risposta individuale, la deviazione campionaria e la distorsione di risposta, rispettivamente come:

$$d_{ijt} = y_{ijt} - m_i \tag{4}$$

$$\Delta_j = m_i - M \tag{5}$$

$$B = M - \mu \tag{6}$$

Sotto queste ipotesi, la media campionaria $\bar{y}_t$ ha un errore totale che, espresso in termini di MSE, risulta pari a:

$$\text{MSE}(\bar{y}_t) = E(\bar{y}_t - \mu)^2 = \sigma^2_{\bar{m}} + \sigma^2_{\bar{d}_t} + 2\,\sigma_{\bar{d},\bar{m}} + B^2 \qquad (7)$$

In presenza di errori di misura, lo stimatore $\bar{y}_t$ è distorto di una quantità pari a B e la sua varianza è maggiore in quanto alla varianza campionaria, $\sigma^2_{\bar{m}}$, si aggiungono la varianza di risposta, $\sigma^2_{\bar{d}_t}$, e la covarianza tra deviazioni campionarie e di risposta $\sigma_{\bar{d},\bar{m}}$, che può assumersi positiva.

La varianza di risposta, in particolare, si può esplicitare nel seguende modo:

$$\sigma^2_{\bar{d}_t} = \frac{1}{n}\,\sigma^2_d + \frac{n-1}{n}\,\varrho\,\sigma^2_d \qquad (8)$$

dove σ^2_d è la varianza di risposta semplice, data da:

$$\sigma^2_d = E\left(\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n'} d^2_{ijt}\right) \qquad (9)$$

e ϱ è il coefficiente di correlazione tra le deviazioni di risposta relative a unità assegnate allo stesso rilevatore e misura, quindi, l'«effetto rilevatore»; in simboli:

$$\varrho = \frac{1}{\sigma^2_d} E\left(\frac{1}{n(n'-1)}\sum_{j=1}^{k}\sum_{i\neq i'}^{n'} d_{ijt}\,d_{i'jt}\right) \qquad (10)$$

La (8) evidenzia l'influenza degli intervistatori sulla varianza di risposta totale; a parità di dimensione campionaria n, essa può essere ridotta diminuendo la numerosità delle assegnazioni di ciascun rilevatore.

Tale modello base è stato successivamente esteso da Fellegi (1964) allo scopo di fornire un quadro di riferimento per l'applicazione congiunta del metodo della reintervista e di quello della compenetrazione del campione che sono generalmente utilizzati per stimare alcune componenti della (7). L'applicazione contemporanea di queste due tecniche rende possibile la stima di un numero maggiore di parametri rispetto al caso in cui se ne utilizzi una sola. In particolare, Fellegi esplicita i coefficienti di correlazione che si possono ot-

tenere combinando differenti tipi di deviazione di risposta e ne fornisce opportuni stimatori. Poiché il modello prevede la replicazione dell'indagine, introduciamo l'indice r = 1 oppure r = 2 per riferirci all'indagine originaria o alla replicazione. Oltre al coefficiente di correlazione tra deviazioni di risposta ottenute da uno stesso intervistatore in una data indagine, ϱ_r, definito dalla (10), i principali coefficienti considerati sono:

— il coefficiente di correlazione tra deviazioni di risposta ottenute da intervistatori diversi nella stessa indagine:

$$\delta_r = \frac{1}{\sigma_d^2} E \left(\frac{1}{nn' (k - 1)} \sum_{j \neq j'}^{k} \sum_{i \neq i'}^{n'} d_{ijr} \, d_{i'j'r} \right) \qquad (11)$$

Tale correlazione è attribuibile all'influenza degli istruttori e dei supervisori comuni a più rilevatori; δ_r misura, quindi l'«effetto supervisore»;

— il coefficiente di correlazione tra deviazioni di risposta relative alla stessa unità nelle due indagini:

$$\beta = \frac{1}{\sigma_{d_1} \sigma_{d_2}} E \left(\frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n'} d_{ij1} \, d_{ij_2 2} \right) \qquad (12)$$

(j_2 indica il rilevatore che ha intervistato l'unità i nella reintervista). Poiché la (12) sarebbe nulla se le due indagini fossero indipendenti, ne consegue che β misura l'«effetto ricordo» della risposta precedentemente data.

Un'estensione del modello di Hansen, Hurwitz e Bershad, che si dimostra particolarmente utile per l'applicazione ad indagini complesse, è stata proposta da Koch (1973). Koch, infatti, ha sviluppato un modello multivariato che può essere adattato a descrivere le componenti dell'errore di misura di stimatori riferiti a disegni di campionamento complessi ([1]). A tale scopo sia u_i una variabile aleatoria indicatrice, tale che:

$$u_i = \begin{cases} 1 \text{ se l'i-esimo elemento della popolazione è nel campione} \\ 0 \text{ altrimenti} \end{cases}$$

Sia ϕ_i la probabilità di selezione dell'i-esimo elemento della popolazione e sia $\phi_{ii'}$ la probabilità di selezione congiunta degli elementi i e i'. Attraverso la distribuzione di probabilità delle variabili u_i è, quindi, possibile descrivere le caratteristiche del disegno campionario. Di conseguenza nelle variabili aleatorie u_i è identificata la fonte di variabilità costituita dagli errori di campionamento, mentre nelle variabili aleatorie y_{it}, ovvero nei valori osservati, quella costituita dagli errori di risposta.

Consideriamo la statistica y_t, ottenuta come combinazione lineare dei valori osservati con pesi noti w_i; ovvero:

$$y_t = \sum_{i=1}^{N} w_i \, u_i \, y_{it} \tag{13}$$

Il valore atteso di risposta individuale, definito dalla (1), si può esprimere come:

$$m_i = E(y_{it} \mid u_i = 1) \tag{14}$$

mentre è conveniente generalizzare la (2) nel seguente modo:

$$M = \sum_{i=1}^{N} w_i \, \phi_i \, m_i \tag{15}$$

Sotto queste ipotesi, la varianza di risposta, σ_r^2, della statistica y_t ha la seguente espressione:

$$\sigma_r^2 = \sum_{i=1}^{N} w_i^2 \phi_i \, E(d_{it}^2 \mid u_i = 1) + \sum_{i \neq i'}^{N} w_i \, w_{i'} \, \phi_{ii'} \, E(d_{it} \, d_{i't} \mid u_i = u_{i'} = 1) \tag{16}$$

dove il primo valore atteso rappresenta la varianza di risposta semplice per l'i-esimo elemento ed il secondo la covarianza tra le deviazioni di risposta relative agli elementi i ed i'. Allora la prima componente della (16) misura la dispersione dei valori osservati dai rispettivi valori attesi, mentre la seconda misura la correlazione tra tali dispersioni per coppie di valori.

In maniera analoga si ottiene un'espressione generalizzata della varianza di campionamento che può essere adattata a disegni campionari complessi mediante la specificazione delle probabilità di selezione ϕ_i e $\phi_{ii'}$.

Nel caso particolare in cui y_t coincida con la media campionaria, equivalente a porre $w_i = 1/(N\phi_i)$, la (16) si riduce ad un'espressione della stessa forma della (8) nella quale la varianza di risposta semplice e la componente correlata hanno espressioni generalizzate per campioni complessi.

Gli stimatori degli effetti degli errori di misura saranno illustrati rispetto al modello base di Hansen, Hurwitz e Bershad, al quale si è fatto riferimento nelle applicazioni effettuate dall'Istat. La formalizzazione adottata consente di utilizzare i seguenti metodi di stima: la reintervista, con o senza riconciliazione delle risposte, e la compenetrazione del campione.

Allo scopo di stimare la distorsione B, (6), è necessario adottare un processo di misurazione più preciso dell'indagine originaria. Si ricorre, quindi, alla reintervista con riconciliazione nella quale il rilevatore è fornito delle risposte originarie e può in caso di discordanza cercare di appurare, con l'aiuto dell'intervistato, quale sia la risposta «vera». Altri accorgimenti per ottenere una misurazione più accurata da assumere quale valore «vero» possono essere, ad esempio, l'utilizzazione di un questionario più dettagliato con domande di controllo e l'impiego di intervistatori più esperti i quali abbiano ricevuto una preparazione migliore ed istruzioni più dettagliate.

Per ottenere una stima della varianza totale e delle sue componenti, si possono utilizzare sia la reintervista sia la compenetrazione del campione. In questo caso la reintervista deve costituire una replicazione indipendente del processo di misurazione, condotta sotto le stesse condizioni essenziali. Si devono, invece, variare le condizioni particolari che si vogliono sottoporre a controllo per valutare la loro influenza sulla qualità dei dati. Per misurare, ad esempio, l'aumento di variabilità delle stime campionarie attribuibile all'impiego dei rilevatori, si devono utilizzare, nella replicazione dell'indagine, intervistatori diversi da quelli originari ma della stessa capacità e che abbiano ricevuto la medesima formazione.

L'«effetto rilevatore» può essere stimato anche mediante la compenetrazione del campione. Questa tecnica si basa sulla suddivisione casuale del campione in tanti sottocampioni di uguale numerosità quanti sono i rilevatori e sull'assegnazione a caso di ciascun sottocampione ad un rilevatore diverso. Poiché i sottocampioni così ottenuti non risultano statisticamente indipendenti, è necessario pianificare l'orga-

nizzazione sul campo in modo tale da eliminare la correlazione tra gli errori di misura di unità appartenenti a sottocampioni diversi, dovuta ad esempio all'«effetto supervisore».

Nel prospetto seguente si riportano gli stimatori ottenibili con ciascun metodo di stima.

	METODO DI STIMA		
COMPONENTE STIMATA	reintervista con riconciliazione	reintervista senza riconciliazione	compenetrazione del campione
distorsione	$\bar{y}_1 - \bar{y}_2$		
varianza totale			$\frac{1}{n} S_b^2$
varianza campionaria		$\frac{1}{n} G$	$\frac{1}{n} S_w^2$
varianza di risposta		$C = \frac{1}{2}(\bar{y}_1 - \bar{y}_2)^2$	
varianza di risposta semplice		$\frac{1}{2n} D$	
componente correlata		$C - \frac{1}{2n} D$	$\frac{n' - 1}{n n'}(S_b^2 - S_w^2)$

dove:

$$G = \frac{1}{2(n-1)} \sum_{r=1}^{2} \sum_{j=1}^{k} \sum_{i=1}^{n'} (y_{ijr} - \bar{y}_r)^2 \qquad (17)$$

$$D = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n'} (y_{ij1} - y_{ij_2 2})^2 \qquad (18)$$

$$S_b^2 = \frac{n'}{(k-1)} \sum_{j=1}^{k} (\bar{y}_{jt} - \bar{y}_t)^2 \qquad (19)$$

$$S_w^2 = \frac{1}{k(n'-1)} \sum_{j=1}^{k} \sum_{i=1}^{n'} (y_{ijt} - \bar{y}_{jt})^2 \qquad (20)$$

Nel confrontare i vari metodi di stima si devono valutare, congiuntamente alle proprietà teoriche degli stimatori, i problemi pratici che sorgono per la loro applicazione alle indagini condotte dall'Istituto, in particolare quelle sulla popolazione. È allora necessario considerare il livello di disaggregazione delle stime, i costi non solo finanziari ma anche organizzativi, il tempo necessario per la raccolta e l'elaborazione dei dati ed infine le risorse disponibili e quindi l'adeguatezza della rete di rilevazione e della organizzazione del lavoro sul campo.

Il presupposto indispensabile per una corretta applicazione di tali metodi ed utilizzazione dei risultati, è l'efficienza dei codici identificativi degli intervistatori e delle unità di rilevazione e di analisi. Questo richiede un maggior controllo del rispetto delle regole di codifica da parte dei comuni e dei rilevatori e l'adeguamento delle procedure che non prevedono l'utilizzazione di alcuni di tali codici.

La reintervista è una tecnica sicuramente costosa e che richiede tempi di attuazione piuttosto lunghi, anche se viene limitata ad un sottocampione. In quest'ultimo caso sorgono ulteriori problemi teorici per l'estensione dei risultati all'intero campione originario. Per le indagini che prevedono la possibilità di risposte «proxy», si deve porre particolare attenzione per accertarsi che il rispondente sia lo stesso in entrambe le indagini, in modo da evitare l'introduzione di distorsioni causate dal cambiamento del rispondente. Infine la stima della distorsione di risposta, mediante la reintervista con riconciliazione, comporta problemi organizzativi e costi aggiuntivi in quanto è necessario utilizzare dei rilevatori particolarmente qualificati.

La compenetrazione del campione non implica costi aggiuntivi per l'indagine, tuttavia richiede una maggiore attenzione ed organizzazione nella fase della rilevazione sul campo. A tale scopo è necessario sensibilizzare ulteriormente i comuni per ottenere una più attenta collaborazione, ed un maggior rispetto delle norme riguardanti le assegnazioni dei rilevatori ed il sistema di codici identificativi. Un problema per l'applicazione del metodo è costituito da quei comuni, rappresentativi dello strato, nei quali il numero di famiglie campione consente una remunerazione «accettabile» ad un solo rilevatore. Si potrebbero, allora, formare degli strati più ampi tali da consentire un numero di interviste più elevato in un solo comune; questo però potrebbe comportare un tasso di campionamento inaccettabile. In alternativa si potrebbero estrarre due comuni per strato e compenetrare le assegnazioni dei due rilevatori su entrambi i comuni anche se ciò comporterebbe un incremento dei costi unitari per intervista.

Le informazioni sugli errori non campionari, così ottenute, possono essere impiegate per il controllo della rete di rilevazione, aspetto di particolare importanza nella determinazione della qualità dell'informazione prodotta. A questo proposito è possibile utilizzare alcuni indicatori desumibili dall'applicazione del metodo della compenetrazione del campione (per il controllo a livello comunale) o della reintervista con riconciliazione (per il controllo a livello rilevatore). Questi indicatori, integrati con altri ottenuti mediante i controlli eseguiti in fasi differenti (ad esempio risultati dei piani di compatibilità), confluendo nell'«archivio rilevatori», consentono di individuare i livelli (comuni, rilevatori, ecc.) per i quali si rende necessario un intervento mirato.

7. I CONTROLLI A REGIME

Il numero ed il tipo dei controlli da effettuare, ovvero il sistema di controllo della singola indagine, dipende dalle risorse disponibili, dai tempi di esecuzione preventivati e dalla rilevanza dei risultati prodotti.

Per ragioni organizzative e di costo, non è generalmente possibile sottoporre la singola rilevazione all'insieme dei controlli su tutte le fasi e sulle possibili fonti di errore. Si può tuttavia formulare un insieme minimo di controlli da prevedere a regime, essenziali per la corretta conduzione dell'indagine ed alcuni facilmente ottenibili come sottoprodotti delle usuali procedure:

- l'analisi del questionario secondo la metodologia «Entità -Relazione» ed il pre-test sul questionario;
- l'analisi del piano di codifica e di registrazione mediante simulazione di un campione di questionari;
- l'analisi del sistema dei codici identificativi;
- l'analisi delle informazioni sulle mancate risposte totali;
- l'analisi delle informazioni desunte dai documenti amministrativi e/o dell'indagine;
- l'analisi delle informazioni desunte dai controllli sulla perforazione;
- la simulazione del piano di compatibilità;
- l'analisi delle informazioni desunte dai piani di compatibilità;
- i controlli sulle tavole prodotte;
- i controlli mediante macro-dati.

NOTE

(*) Il contributo degli autori alla presente relazione si suddivide nel seguente modo: M. Masselli dal paragrafo 1 al paragrafo 5 e paragrafo 7, M. Signore per il paragrafo 6.

(1) Per omogeneità con i modelli già illustrati, tralasceremo l'estensione multivariata proposta da Koch; inoltre ometteremo l'indice j riferito, in precedenza, al rilevatore.

RIFERIMENTI BIBLIOGRAFICI

COCHRAN W. (1977), *Sampling Techniques*, J. Wiley, New York.

FELLEGI I. (1964), «*Response Variance and Its Estimation*», Jour. Amer. Stat. Assoc., Vol. I, pp. 1016-1041.

FELLEGI I.P., HOLT D. (1976), «*A systematic approach to automatic editing & imputation*», J.A.S.A., pp. 17-35.

GARCIA RUBIO, GOMEZ ALONSO, VILLAN (1983), «*Desarollo de un sistema de detection y imputation automatica basando en la metodologia de Fellegi-Holt ampliada*», Atti I.S.I., Vol. I, pp. 54-58.

HANSEN M.H., W.N. HURWITZ e M.A. BERSHAD (1961), «*Measurement Errors in Censuses and Surveys*», Bull. Int. Stat. Inst., n. 38, Vol. II, pp. 359-374.

I.N.S.E.E. (1985), «*Rapport sur la qualité des travaux statistiques*», Parigi, documento interno.

KOCH G G (1973), «*An alternative approach to multivariate response error models with applications to estimators involving subclass means*», J.A.S.A., Vol. 68, pp. 906-913.

LITTLE R.J.A., RUBIN D.B. (1987) «*Statistical analysis with missing data*», J. Wiley & Sons, New York.

MADOW W.G., OLKIN I., RUBIN D.B. (1983), «*Incomplete data in sample surveys*», Academic Press, New York.

Statistics Canada (1976), «*A compendium of methods of error evaluation in censuses and surveys*».

T. WRIGTH (1983), «*Statistical methods and the improvement of data quality*», Academic Press, New York.

U.N. (1982), «*National household survey capability programme. Non-sampling errors in household surveys: sources, assesment and control*», New York.

CONTRIBUTI DEL PROGETTO QUALITÀ DEI DATI ED ALTRI CONTRIBUTI ISTAT

ABBATE C.C. (1989), «*Indagini sperimentali nel quadro della ristrutturazione del sistema di indagini agricole in Italia - Controllo della qualità dei dati*», documento interno.

BARCAROLI G., FORTUNATO E., MAGALOTTI, MANICARDI G., VACCARI C. (1987), «*Manuale per la progettazione concettuale di dati statistici*», Istat, Roma.

CORTESE A. (1983), «*Indagine sul confronto censimento-anagrafe: scopi, modalità d'esecuzione, principali risultati*», Atti del Convegno S.I.S., Trieste, pp. 121-144.

DE MARCHIS M.A. (1988), «*Interviewer file of Istat household surveys*», Conferenza I.A.O.S., Roma, pp. 112-115.

MAGANO S. (1984), «*Analisi dell'influenza dei rilevatori sulla qualità dei dati raccolti nel terzo censimento generale dell'agricoltura, attraverso il metodo dell'analisi della varianza*», Atti della XXXII Riunione scientifica della S.I.S., Sorrento, Vol. I, pp. 421-430.

MANICARDI G., VENTURI M. (1988), «*Analisi integrata di dati e funzioni nei sistemi informativi statistici*», documento interno.

MARCHETTI E. (1986), «*Large sample models for editing response errors*», documento interno.

MASSELLI M. (1983), «*Risultati dell'indagine di controllo sulla qualità dei dati del censimento 1981*», Atti del Convegno S.I.S., Trieste, pp. 145-169.

MASSELLI M. (1985), «*La qualità dei dati nelle rilevazioni statistiche*», Rivista Italiana di Economia, Demografia e Statistica, Vol. 40.

MASSELLI M. (1985), «*Nota sul progetto qualità dei dati*», documento interno.

MASSELLI M. (1986), «*Valutazione dei piani di compatibilità e correzione automatici. Una sperimentazione*», Atti della XXXIV Riunione Scientifica della S.I.S., Bari, Vol. II, pp. 257-264.

MASSELLI M. (1987), «*La procedura di controllo degli effetti del piano di compatibilità dell'indagine forze di lavoro*», documento interno.

MASSELLI M. (1988), «*La procedura di controllo dell'indagine forze di lavoro*», documento interno.

MASSELLI M. (1988), «*L'errore di identificazione delle unità ed il sistema di controllo di un'indagine statistica, Una applicazione all'indagine sulle forze di lavoro*», Atti della XXXIV Riunione Scientifica della SIS, Siena, Vol. II, Tomo I, pp. 169-176.

MASSELLI M., DI PIETRO E., DE MARCHIS M.A., SIGNORE M. (1988), «*Obiettivi e metodi di controllo dell'indagine pilota - Indagine sulla storia lavorativa*», documento interno.

MASSELLI M., DI PIETRO E., PANIZON F., SIGNORE M. (1986), «*Il sistema di codifica e la ricostruzione longitudinale delle famiglie nell'indagine forze di lavoro*», documento interno.

MASSELLI M., MARCHETTI E. (1984), «*I piani di compatibilità ed il controllo dell'attendibilità del dato*», Atti della XXXII Riunione Scientifica della S.I.S., Sorrento, Vol. I, pp. 439-450.

MASSELLI M., PANIZON F., SIGNORE M. (1988), «*Il sistema di controllo della qualità dei dati*», Manuale di tecniche d'indagine, stesura provvisoria.

MASSELLI M., SIGNORE M. (1989), «*Exact linkage problems in the Italian labour force survey*», Proceedings of the 47th session of I.S.I., Parigi.

MASSELLI M., TERRA ABRAMI V. (1983), «*L'indagine di controllo di copertura del censimento della popolazione*», Atti del Convegno S.I.S., Trieste, Vol. I, pp. 171-188.

PANIZON F. (1988), «*Il controllo statistico di qualità della fase della registrazione dei dati*», Atti della XXXIV Riunione Scientifica della SIS, Siena, Vol. II, Tomo I, pp. 185-192.

PANIZON F., SIGNORE M. (1987), «*Analisi dell'effetto dei piani di compatibilità dell'indagine forze di lavoro con accoppiamento statistico dei records*», documento interno.

QUINTANO C., CALZARONI M., DINI P., MASSELLI M., POLITI M., TACCINI P. (1987), «*Una ricognizione dell'error profile dell'indagine sul prodotto lordo*» in «Attendibilità e tempestività delle stime di contabilità nazionale», a cura di U. Trivellato, CLEUP Padova.

SCHIRINZI G., (1986), «*Alcune prime annotazioni sulla ripartizione delle aziende agricole secondo la superficie*», Convegno della SIS su «Statistica e risorse naturali», Messina.

SIGNORE M. (1988), «*Evaluation of the Interviewer's Influence on the Quality of the 1985 Sports and Holidays Survey Data*», Pre-Proceedings of the First Conference of IAOS, Roma, pp. 252-256.

SIGNORE M. (1988), «*Stima dell'errore di misura: alcune riflessioni sui problemi teorici e pratici per l'applicazione ad indagini su larga scala*», Atti della XXXIV Riunione Scientifica della SIS, Siena, Vol. II, Tomo I, pp. 193-200.

ZANNELLA F., SABBADINI L.L., BURATTA V., (1986) «*Analisi dell'effetto proxy in alcune recenti indagini sulle famiglie condotte dall'Istat: primi risultati*», documento interno.

ZANNELLA F., SABBADINI L.L., BURATTA V., (1986), «*Analisi dell'effetto proxy nell'indagine sulle forze di lavoro del luglio 1986 - Risultati preliminari*», documento interno.

ZUCHEGNA A. (1984), «*La digitazione dei dati ed il controllo statistico*», Tesi di laurea, Facoltà di Scienze Statistiche, Università di Roma.

L'INDAGINE ISTAT SULLE FORZE DI LAVORO IN UMBRIA: UNA ANALISI EMPIRICA DEL DISEGNO

di *Giorgio Eduardo Montanari*

1. INTRODUZIONE

L'indagine campionaria trimestrale sulle Forze di Lavoro (nel seguito FL) condotta dall'Istituto Centrale di Statistica (Istat) rappresenta una delle più importanti fonti di informazione sulle principali dinamiche dell'offerta di lavoro e dell'occupazione in Italia. Essa permette anche un aggiornamento costante dei dati sulle caratteristiche anagrafiche e professionali della popolazione italiana negli intervalli decennali tra un censimento demografico e l'altro. La sua importanza è poi notevolmente aumentata da quanto il fenomeno della disoccupazione ha assunto dimensioni allarmanti. Si giustifica così la richiesta sempre più pressante di informazioni ad un livello territoriale sempre più ristretto. Non a caso il campione di «base» (Istat, 1978) è stato sovradimensionato in diverse regioni per conoscere nel dettaglio anche la realtà subregionale.

Dal luglio 1988, anche la regione Umbria dispone di un campione ampliato. In funzione di quest'ultimo ampliamento è stato effettuato lo studio empirico, di cui tratta la presente nota, con l'obiettivo di valutare l'efficienza del disegno di campionamento FL e suggerire le modalità del suo ampliamento. Il lavoro svolto ha fornito anche l'occasione di studiare per via empirica come le diverse componenti del piano di campionamento agiscono sul cosiddetto «effetto del disegno» (*design effect*, Kish, 1965). Riteniamo che i risultati ottenuti siano interessanti per le numerose indicazioni che suggeriscono.

Per la descrizione del disegno di campionamento si rimanda a Istat, 1978.

Per quanto riguarda gli stimatori utilizzati, introduciamo la seguente simbologia. Sia h = indice di strato (h = 1,2...,H); i = indice di comune (i = 1,2,...,N_h); j = indice di famiglia (j = 1,2,...,M_{hi}); a = indice di sesso (a = 1,2; 1 = maschio, 2 = femmina); M_h = numero delle famiglie nello strato h; n_h = numero dei comuni campione nello strato h; m_{hi} = numero delle famiglie da intervistare nel comune campione i dello strato h.

Siano poi X_{ahij} = numero componenti di sesso a della famiglia j del comune i dello stato h; Y_{ahij} = numero dei componenti di sesso a che possiedono il carattere in considerazione nella famiglia j del comune i dello strato h. Per queste due variabili useremo qui la convenzione secondo cui l'assenza di uno o più indici sta a significare che si è effettuata la somma rispetto agli indici mancanti (ad esempio X rappresenterà la popolazione complessiva). Sia infine $P_{ah} = X_{ah}/X_h$ il tasso di mascolinità o femminilità nello strato h.

In questa nota prenderemo in considerazione la stima del totale delle unità, Y, che possiedono il carattere oggetto di studio. Queste quantità sono la base per la stima di rapporti fra totali, quali ad esempio i tassi di occupazione e di disoccupazione, la percentuale di popolazione attiva addetta ai vari settori di attività economica, ecc.

Lo stimatore di Y utilizzato nella indagine FL è di tipo per quoziente, separato e post-stratificato rispetto al sesso (nel seguito la post-stratificazione si intenderà sempre rispetto al sesso).

Esso può essere scritto

$$\hat{Y} = \sum_{a=1}^{2} \hat{Y}_a, \qquad (1)$$

dove $\hat{Y}_a$ è lo stimatore del totale per il sesso a, la cui espressione, indicando con

$$\hat{R}_{ahi} = \frac{\sum_{j=1}^{m_{hi}} Y_{ahij}}{\sum_{j=1}^{m_{hi}} X_{ahij}}$$

lo stimatore di $R_{ahi} = Y_{ahi}/X_{ahi}$ nell'unico comune campione i dello strato h, è dato da

$$\hat{Y}_a = \sum_h \hat{R}_{ahi} X_{ah}. \qquad (2)$$

È da notare che lo stimatore (1) non tiene conto della struttura rotante del campione: esso opera come se la rilevazione fosse di tipo *cross-section*.

In vista dei confronti da effettuare nel seguito, è opportuno estendere la (2) al caso di più comuni campione per strato. Questo può

essere fatto in due modi diversi, a seconda che si opti per uno stimatore del tipo media di rapporti o del tipo rapporto di medie. Nel primo caso si ha

$$\hat{Y}_a = \sum_h \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{R}_{ahi} X_{ah} \qquad (3)$$

e nel secondo

$$\hat{Y}_a = \sum_h \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} Y_{hij}}{\sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} X_{hij}} X_{ah},$$

dove w_{hij} è l'inverso della probabilità di inclusione nel campione della famiglia j del comune i dello strato h. Tale probabilità, che indicheremo con π_{hij}, è data da

$$\pi_{hij} = \pi_{hi} \frac{m_{hi}}{M_{hi}},$$

dove π_{hi} è la probabilità di inclusione del comune i. Nel seguito assumeremo che π_{hi} soddisfi la relazione

$$\pi_{hi} = n_h X_{hi}/X_h = n_h z_{hi}.$$

Tra le due possibili generalizzazioni, per esigenze di calcolo nello studio empirico che seguirà, abbiamo adottato lo stimatore (3).

Lo stimatore $\hat{Y}$ con $\hat{Y}_a$ dato dalla (3) risulta approssimativamente corretto. Per quanto riguarda la varianza, allo scopo di simulare il campionamento senza ripetizione dei comuni (che sono le unità di primo stadio, UPS, mentre le unità di secondo stadio sono le famiglie) senza ricorrere ad uno dei numerosissimi metodi proposti allo scopo in letteratura, si moltiplicherà la varianza propria del campionamento con ripetizione per il fattore $(N_h - n_h)/(N_h - 1)$ (cfr. Fellegi et al., 1967; Gray, 1975), non influente se si estrae un solo comune per strato. Ciò per evitare di appesantire oltre misura il lavoro. Si ottiene, per campioni sufficientemente grandi,

$$V(\hat{Y}) \doteq \sum_h \frac{N_h - n_h}{N_h - 1} \frac{1}{n_h} \sum_{i=1}^{N_h} z_{hi} \Big(\sum_{a=1}^{2} R_{ahi} X_{ah} - \sum_{i=1}^{N_h} z_{hi} \sum_{a=1}^{2} R_{ahi} X_{ah} \Big)^2 +$$

$$+ \sum_h \frac{1}{n_h} \sum_{i=1}^{N_h} z_{hi} V_2 \Big(\sum_{a=1}^{2} \hat{R}_{ahi} X_{ah} \Big), \qquad (4)$$

dove V_2 denota l'operatore di varianza dato il comune i. Il primo ed il secondo termine dopo il segno di uguaglianza nella (4) sono rispettivamente le varianze di primo stadio (VPS) e di secondo stadio (VSS).

Autoponderazione, stratificazione e tasso di sondaggio. Un campione si dice autoponderato quando tutte le unità della popolazione hanno una probabilità di inclusione pari al tasso di sondaggio f. Perché ciò si verifichi occorre che sia soddisfatta, per ogni h ed i, la relazione

$$m_{hi} = f \frac{X_h M_{hi}}{n_h X_{hi}}.$$

Nella indagine FL, per semplicità, si sostituisce il rapporto M_{hi}/X_{hi} con M_h/X_h (queste sono delle quantità praticamente costanti) ottenendo

$$m_{hi} = m_h = fM_h/n_h. \tag{5}$$

Quando il campione è autoponderato si può stabilire un interessante risultato concernente l'effetto della stratificazione e del tasso di sondaggio su VPS e VSS. Sotto ipotesi piuttosto generali si dimostra che, ponendo $V_2 (\Sigma_{a=1}^{2} \hat{R}_{ahi} X_{ah}) = m_{hi} X_h^2 V_U (\Sigma_{a=1}^{2} \hat{R}_{ahi} P_{ah})$, la (4) si può scrivere

$$V(\hat{Y}) = \sum_h \frac{N_h - n_h}{N_h - 1} \frac{1}{n_h} \sum_{i=1}^{N_h} z_{hi} \left(\sum_{a=1}^{2} R_{ahi} X_{ah} - \sum_{i=1}^{N_h} z_{hi} \sum_{a=1}^{2} R_{ahi} X_{ah} \right)^2 +$$

$$+ f^{-1} \sum_h \sum_{i=1}^{N_h} \frac{X_h}{M_h} X_{hi} V_U \left(\sum_{a=1}^{2} \hat{R}_{ahi} P_{ah} \right) +$$

$$- \sum_h \sum_{i=1}^{N_h} \frac{X_h}{n_h} \frac{X_{hi}}{M_{hi}} V_U \left(\sum_{a=1}^{2} \hat{R}_{ahi} P_{ah} \right). \tag{6}$$

Dalla (6) si evince che VPS dipende esclusivamente dalla stratificazione dei comuni e dalla allocazione delle unità di primo stadio campione (UPSC) tra gli strati. Per quanto riguarda VSS il termine nella seconda riga della (6) rappresenta VSS al lordo della correzione per popolazione finita. Esso dipende dal tasso di sondaggio f, è indipendente dal numero delle UPSC ed è praticamente indipendente dalla stratificazione, dal momento che le quantità che compaio-

no a livello di strato — cioè il rapporto X_h/M_h, pari al numero medio di individui per famiglia, ed il tasso di mascolinità o femminilità P_{ah} — sono approssimativamente costanti essendo delle caratteristiche demografiche. L'ultimo termine della (6), di entità generalmente piccola, rappresenta la correzione per popolazione finita. Esso è indipendente dal tasso di sondaggio ma non dal numero delle UPSC.

In sintesi, VSS dipende sostanzialmente dal tasso di sondaggio ed è indipendente dalla struttura del campione di primo stadio se si eccettua il termine di correzione per popolazione finita. Un analogo risultato è stato mostrato da Zannella (1987) in un contesto meno complesso di quello da noi considerato.

2. UNO STUDIO EMPIRICO

Lo scopo del lavoro è quello di valutare l'efficienza del disegno di campionamento per le FL simulandone l'applicazione sulla popolazione reale, interamente nota, costituita dai dati individuali del censimento demografico 1981 in Umbria (escluse le convivenze). I dati censuari, infatti, sono una fotografia, se così si può dire, della popolazione oggetto dell'indagine FL e, mediante le informazioni riguardanti le condizioni lavorative delle persone censite, permettono di «riempire parzialmente il questionario» delle FL. Si può così stabilire l'appartenenza degli individui ai diversi aggregati. In questo studio sono stati considerati quelli di maggiore interesse ai fini della conoscenza del mercato del lavoro, precisamente essi sono:

— addetti all'agricoltura (AGR);
— addetti all'industria (IND);
— addetti alle altre attività (ATT);
— totale occupati (OCC = AGR + IND + ATT);
— in cerca di nuova occupazione (NOC);
— in cerca di prima occupazione (POC);
— totale disoccupati (DIS = NOC + POC);
— forze di lavoro (FLA = OCC + DIS).

L'assegnazione delle persone ai vari aggregati è stata fatta in base alle definizioni adottate nell'indagine sulle FL, compatibilmente con le informazioni contenute nei dati censuari.

Lo schema di campionamento per le FL in Umbria prima del recente ampliamento prevedeva la selezione di 25 comuni (9 con popolazione superiore alle 20.000 unità e 16 tra i rimanenti comuni) su un totale di 92. Il tasso di sondaggio, f, pari allo 0,6% era tale che ad ogni rilevazione la raccolta dei dati coinvolgeva circa 1.600 famiglie (approssimativamente 4800 persone).

Lo studio empirico è stato impostato nel modo seguente. Dal momento che il nastro dei dati censuari è organizzato per comune ed all'interno di questo l'ordinamento delle famiglie rispecchia quello della rispettiva anagrafe — che è organizzata per sezioni di censimento — si è proceduto alla estrazione, da ciascun comune, di tutti i campioni di famiglie sistematici aventi la dimensione prescritta, m_h (determinata ponendo $f = 0{,}006$ nella (5) e con riferimento alla stratificazione per le FL) ed al calcolo delle stime $\hat{R}_{1hi}$ ed $\hat{R}_{2hi}$ per gli otto aggregati considerati. Ciò consente di tenere conto del meccanismo di formazione del campione il quale garantisce che la distribuzione sul territorio comunale delle famiglie estratte rifletta quella dell'universo in modo proporzionale (cfr. Istat, 1978; pagg. 29-30). Successivamente sono stati determinati, per ciascun aggregato, il valore atteso, la varianza, nonché la covarianza di $\hat{R}_{1hi}$ ed $\hat{R}_{2hi}$ nell'universo dei campioni estratti in ciascun comune e quindi i valori esatti del valore atteso e della varianza di $\hat{Y}_a$ e $\hat{Y}$.

I risultati concernenti la distorsione, il coefficiente di variazione, il rapporto tra VPS e VSS e l'effetto del disegno (sinteticamente *deff*) degli stimatori del totale sono riportati nella tabella 1 dove il totale del carattere nella popolazione è espresso in percentuale. Tali grandezze sono distinte per sesso e per provincia.

Dall'analisi della tabella seguono alcune indicazioni interessanti.

Distorsione. La distorsione relativa degli stimatori utilizzati nella indagine FL è sempre piuttosto contenuta ed inferiore al mezzo punto percentuale per tutte le stime regionali. Si osservi poi che in provincia di Terni la distorsione è in modulo mediamente più elevata rispetto alla provincia di Perugia (ciò può essere imputato, almeno in parte, al diverso numero medio di famiglie campione per comune: 51 a Terni contro un valore di 70 in provincia di Perugia) mentre la distorsione degli stimatori di POC e DIS, oltre ad essere tra quelle più cospicue, è sempre di segno negativo.

Precisione delle stime. La precisione delle stime è misurata dai rispettivi coefficienti di variazione: tanto più questi sono grandi, tanto più la precisione è bassa. Osserviamo prima di tutto che i coefficienti sono più grandi per gli aggregati più piccoli, come era da attendersi avendo a che fare con popolazioni bernoulliane; inoltre essi differiscono nelle due provincie a causa della diversa consistenza del campione (circa 460 famiglie in provincia di Terni e 1.130 in quella di Perugia). È da notare il valore per il totale regionale dell'aggregato FLA, ben al di sotto di quel 5% fissato in Istat (1978) per il dimensionamento del campione a livello regionale. Tuttavia occorre precisare che nella realtà i coefficienti di variazione effettivi possono risultare più elevati, a causa degli errori non campionari di cui non si è tenuto conto.

Varianze di primo e secondo stadio. Particolarmente interessante è il risultato concernente il rapporto tra le VPS e VSS: il contributo di VPS alla varianza totale risulta molto basso (un analogo risultato è stato mostrato per quanto riguarda il Current Population Survey degli Stati Uniti; Hanson, 1978). Si passa da un massimo del 37% di AGR ad un minimo del 3% per POC con riferimento al totale in Umbria. La determinante che maggiormente spiega questo fenomeno è la presenza dei comuni autorappresentativi, che non contribuiscono alla varianza di primo stadio. Essi contengono in provincia di Perugia circa il 56% della popolazione e in quella di Terni addirittura il 69% (infatti in quest'ultima provincia la frazione di varianza di primo stadio è sistematicamente inferiore).

Design effect. In provincia di Terni i valori dell'effetto del disegno sono sistematicamente più bassi rispetto a quelli della provincia di Perugia e qui valgono le considerazioni già fatte per la varianza di primo e di secondo stadio. Per l'intera regione i valori più elevati sono quelli di AGR e IND mentre i più bassi sono quelli relativi a OCC e FLA soprattutto tra i maschi. È da osservare come l'effetto del disegno cambi non solo al variare dell'aggregato considerato, come è stato osservato da O'Muircheartaigh (1978) e Fabbris (1981), ma anche del dominio territoriale. Pertanto la tabella fornita dall'Istat (1978; pag. 61) per il calcolo dello scarto teorico assoluto, essendo basata sul presupposto che la varianza delle stime dipende solo dalla grandezza dell'aggregato, produce valori con un basso livello di approssimazione.

Tabella 1 - Percentuale nella popolazione (a) di alcuni aggregati di interesse in Umbria e nelle provincie di Perugia e Terni (M = maschi; F = femmine) e distorsione percentuale (b), coefficiente di variazione percentuale (c), frazione su quella totale della varianza di primo stadio (d) ed effetto del disegno (e) degli stimatori $\hat{Y}_a$ e $\hat{Y}$.

AGGREGATO		PERUGIA			TERNI			UMBRIA		
		M	F	Tot.	M	F	Tot.	M	F	Tot.
	a	6,91	2,03	4,43	4,94	0,93	2,89	6,35	1,72	3,99
	b	—0,09	0,15	—0,04	—0,08	0,48	—0,19	—0,03	0,21	—0,07
AGR	c	11,79	19,72	11,64	19,56	44,07	19,25	10,16	18,00	10,07
	d	0,36	0,30	0,40	0,18	0,15	0,21	0,33	0,28	0,37
	e	1,75	1,42	2,17	1,31	1,31	1,50	1,65	1,40	2,03
	a	22,26	9,75	15,90	25,06	3,20	13,89	23,05	7,90	15,33
	b	0,08	—0,08	0,03	0,29	—0,52	0,17	0,12	—0,11	0,06
IND	c	4,79	8,02	4,50	6,85	21,94	6,63	3,93	7,53	3,75
	d	0,15	0,25	0,23	0,08	0,15	0,09	0,13	0,24	0,20
	e	1,13	1,23	1,33	1,05	1,12	0,97	1,10	1,20	1,24
	a	21,32	13,80	17,50	18,47	13,60	15,98	20,52	13,74	17,07
	b	0,20	0,49	0,32	0,28	—0,19	0,10	0,22	0,34	0,27
ATT	c	4,62	6,07	4,08	8,19	9,65	6,83	4,03	5,14	3,50
	d	0,13	0,06	0,14	0,03	0,03	0,03	0,11	0,05	0,11
	e	0,99	1,05	1,23	1,00	1,01	1,20	0,99	1,04	1,22
	a	50,49	25,58	37,83	48,47	17,73	32,76	49,92	23,36	36,40
	b	0,09	0,21	0,14	0,14	—0,20	0,05	0,10	0,14	0,11
OCC	c	2,03	4,18	2,08	3,31	8,18	3,39	1,73	3,72	1,78
	d	0,06	0,19	0,18	0,03	0,04	0,05	0,06	0,16	0,15
	e	0,72	1,06	0,92	0,68	1,00	0,76	0,71	1,04	0,88
	a	0,92	0,82	0,87	1,18	1,04	1,11	0,99	0,88	0,94
	b	—0,22	0,51	0,12	0,17	—0,71	—0,28	—0,10	0,12	0,01
NOC	c	27,31	26,87	20,42	36,03	40,57	27,16	21,77	22,45	16,33
	d	0,06	0,12	0,13	0,01	0,02	0,02	0,04	0,08	0,10
	e	1,17	1,04	1,25	1,10	1,20	1,17	1,14	1,10	1,22
	a	2,64	3,30	2,98	2,67	3,66	3,18	2,65	3,40	3,03
	b	—0,35	—0,29	—0,35	—0,98	—0,76	—0,87	—0,51	—0,44	—0,45
POC	c	15,51	12,69	10,33	22,65	19,50	14,86	12,83	10,64	8,50
	d	0,02	0,03	0,03	0,01	0,01	0,01	0,02	0,03	0,03
	e	1,10	0,96	1,13	0,95	1,00	0,99	1,06	0,97	1,09
	a	3,56	4,12	3,85	3,85	4,70	4,29	3,64	4,29	3,97
	b	—0,31	—0,13	—0,24	—0,58	—0,77	—0,68	—0,37	—0,30	—0,34
DIS	c	13,38	11,23	9,11	19,29	17,75	13,24	11,01	9,51	7,51
	d	0,04	0,05	0,06	0,01	0,01	0,01	0,03	0,04	0,05
	e	1,12	0,95	1,14	1,02	1,08	1,08	1,08	0,99	1,12
	a	54,02	29,73	41,67	52,33	22,44	37,05	53,55	27,67	40,37
	b	0,07	0,01	0,05	0,09	—0,31	—0,03	0,07	—0,06	0,03
FLA	c	1,91	3,73	1,93	3,19	6,74	3,10	1,64	3,27	1,65
	d	0,05	0,18	0,16	0,04	0,04	0,04	0,05	0,14	0,13
	e	0,73	1,04	0,93	0,74	0,92	0,80	0,73	1,01	0,89

2.1. Scomposizione dell'effetto del disegno

Diamo ora conto dei risultati conseguiti in merito all'impatto delle diverse componenti del disegno sulla efficienza delle stime. Precisiamo che nel seguito i valori di $V_2(\Sigma_a^2 \hat{R}_{ahi} X_{ah})$, quando m_{hi} è diverso da quello sopra definito, sono stati imputati ipotizzando una relazione inversamente proporzionale tra $V_2(\Sigma_a^2 \hat{R}_{ahi} X_{ah})$ e m_{hi}.

Tabella 2 - Varianza di primo stadio (posta pari a 100 nel caso di ST1) e, fra parentesi, frazione sul totale della varianza di primo stadio dello stimatore del totale sotto stratificazioni diverse

AGGREGATO	STRATIFICAZIONE			
	ST1	ST2	ST3	ST4
AGR	100,0	31,0	28,0	22,2
	(0,69)	(0,40)	(0,37)	(0,32)
IND	100,0	48,5	19,9	1,6
	(0,57)	(0,38)	(0,20)	0,02)
ATT	100,0	8,9	8,4	6,5
	(0,61)	(0,12)	(0,11)	(0,09)
OCC	100,0	42,4	25,0	12,0
	(0,42)	(0,23)	(0,15)	(0,08)
NOC	100,0	70,1	76,7	46,7
	(0,13)	(0,09)	(0,10)	(0,06)
POC	100,0	30,7	25,3	20,0
	(0,11)	(0,03)	(0,03)	(0,02)
DIS	100,0	35,5	38,3	27,1
	(0,12)	(0,04)	(0,05)	(0,03)
FLA	100,0	44,3	24,1	13,2
	(0,39)	(0,21)	(0,13)	(0,08)

Ruolo della stratificazione. La stratificazione, come è stato detto in precedenza, agisce sostanzialmente su VPS. Per valutare l'efficacia di quella utilizzata per l'indagine FL è sufficiente calcolare la varianza di primo stadio e quella totale in corrispondenza di alcune stratificazioni alternative. In particolare abbiamo messo a confronto le seguenti stratificazioni:

ST1 - costituita da un unico strato di comuni per provincia;

ST2 - ottenuta dalla precedente rendendo autorappresentativi i comuni con più di 20.000 abitanti;

ST3 - la stratificazione delle FL;

ST4 - ottenuta da ST2 stratificando i comuni non autorappresentativi mettendo insieme quelli che occupano posizioni contigue nel-

la graduatoria secondo la percentuale di popolazione appartenente ad IND ed in modo da costituire strati della stessa dimensione in termini di popolazione.

In tutti i casi le UPSC sono 25 e la loro distribuzione tra le due provincie è quella implicata da ST3; sotto ST3 ed ST4 le UPSC sono una per strato.

Nella tabella 2 sono riportati i valori di VPS (posti pari a 100 nel caso di ST1) e, fra parentesi, la frazione sul totale della varianza di primo stadio corrispondente alle quattro stratificazioni. È da aggiungere che VSS si è mantenuta approssimativamente costante con variazioni percentuali rispetto ai valori di ST3 contenute nell'intervallo (—4,3 - —2,0) per ST1 e (0,0 - 0,9) per ST2 e ST4.

Dall'esame della tabella emergono le considerazioni che seguono. Rispetto a ST1, la stratificazione ST3 adottata nella indagine FL comporta una forte riduzione di VPS (ad eccezione di NOC). Tuttavia, se si osserva la colonna relativa a ST2, se ne deduce che tale riduzione è da attribuire in gran parte alla presenza dei comuni autorappresentativi che da soli raccolgono circa il 60% della popolazione umbra. La percentuale di riduzione di VPS da attribuire a questi ultimi è per gli otto aggregati, nell'ordine, 96%, 64%, 99%, 77%, 100%, 93%, 100% e 73%. L'efficacia della stratificazione dei comuni non autorappresentativi è quindi di minor rilievo. Interessanti sono anche i risultati concernenti ST4 che non tiene conto del settore statistico e della zona altimetrica. Per tutti gli aggregati si conseguono apprezzabili riduzioni di VPS.

Per quanto riguarda l'effetto della stratificazione sulla varianza totale, esso dipende dalla frazione di varianza di primo stadio che nel seguito denoteremo con il simbolo α. Ricordando che la stratificazione lascia sostanzialmente immutata VSS, è facile vedere che passando da una stratificazione ad un'altra, la variazione percentuale della varianza totale è approssimativamente pari a quella di VPS moltiplicata per il valore di α della stratificazione di partenza. Ad esempio, con riferimento alle variabili relative alla disoccupazione, una stratificazione che rispetto a ST1 comportasse una riduzione di VPS del 90% nella migliore delle ipotesi farebbe diminuire la varianza totale di una percentuale non superiore al 12%. L'effetto della stratificazione sulle variabili relative alla disoccupazione è perciò assai limitato. La conclusione dipende ovviamente dal tasso di sondaggio che determina il valore di α. Si può dimostrare che aumentando di

k volte il tasso di sondaggio, ignorando la correzione per popolazione finita al secondo stadio, si avrebbe un incremento di α pari a

$$(1 + \frac{1 - \alpha}{k\alpha})^{-1}.$$

Riprendendo l'esempio precedente e relativamente alla stratificazione ST1, se se si raddoppiasse il tasso di sondaggio (k = 2) il valore di α per AGR salirebbe a 1,29 mentre quello di POC salirebbe soltanto allo 0,19. Pertanto, indipendentemente dal tasso di sondaggio, le variabili relative alla disoccupazione sono quelle che traggono il minor beneficio da una qualsivoglia stratificazione.

Cluster comune. Per valutare l'effetto dell'adozione del comune come UPS, faremo ricorso al campionamento casuale semplice degli individui (CCSI) ed allo stimatore non post-stratificato

$$\hat{Y} = \sum_h \left(\frac{1}{n_h} \sum_{i=1}^{n_h} \hat{R}_{hi}\right) X_h, \qquad (7)$$

rapportandone la varianza nell'ambito del campionamento a due stadi (avente i comuni per UPS e i singoli individui selezionati con il CCSI per unità di secondo stadio) con quella del CCSI all'interno delle province assumendo sempre la condizione di autoponderazione (5). Entrambe le varianze dipendono da parametri della popolazione che sono facilmente calcolabili. Nella tabella 3, seconda colonna, sono riportati tali rapporti (*deff*). Da essi si possono ricavare i valori approssimati delle misure di omogeneità, ϱ, degli individui dentro i comuni, sulla base della espressione sintetica (Kish, 1965)

$$deff = 1 + \varrho\,(x/n - 1),$$

Tabella 3 - Effetto del cluster comune e relativi coefficienti di correlazione intracluster ϱ ed effetto della post-stratificazione e del campionamento sistematico delle famiglie (relativamente a VSS) sulla varianza dello stimatore del totale di alcuni aggregati

AGGREGATO	PSU Comune	ϱ	post-strat.	CSSF
AGR	3,67	0,01389	0,974	1,35
IND	2,24	0,00645	0,950	1,06
ATT	2,63	0,00848	0,991	1,12
OCC	1,51	0,00265	0,922	0,82
NOC	1,14	0,00073	0,999	1,12
POC	1,11	0,00057	0,999	1,07
DIS	1,12	0,00062	0,999	1,08
FLA	1,46	0,00239	0,928	0,85

dove x è l'ampiezza del campione in termini di individui ed n è il numero delle UPSC. Si osservi l'effetto inflattivo piuttosto consistente del cluster comune sugli addetti ai tre settori di attività economica e, al contrario, l'effetto trascurabile sulle variabili relative alla disoccupazione.

Post-stratificazione. Si noti che la post-stratificazione agisce su VSS lasciando praticamente inalterata VPS. Per valutare perciò l'effetto della post-stratificazione sulla varianza di secondo stadio, assumendo il disegno di campionamento a due stadi descritto immediatamente dopo la (7), occorre determinare le VSS degli stimatori (1) e (7) e farne il rapporto. Detti rapporti sono riportati nella tabella 3, quarta colonna. È evidente che la post-stratificazione è tanto più efficace quanto più sono diverse le percentuali dei due sessi tra gli appartenenti agli aggregati considerati. Il maggior guadagno si ha per IND (5,0%), OCC (7,8%), FLA (7,2%) ed in misura minore per AGR (2,6%); è presso irrilevante per le variabili relative alla disoccupazione.

Campionamento sistematico delle famiglie. Lo studio empirico precedentemente descritto ha permesso di calcolare i valori di VSS dello stimatore (1) nell'ambito del campionamento sistematico delle famiglie (CSSF) nei comuni. Rapportando tali valori con quelli calcolati assumendo il CCSI dentro i comuni, è possibile ora valutare l'effetto del CSSF. I rapporti ottenuti nell'ipotesi della stratificazione ST1 sono riportati nella tabella 3, ultima colonna. Si osservino i valori elevati di AGR, ALT e NOC che denotano consistenti aumenti di VSS. Traggono invece beneficio dal CSSF le variabili OCC e FLA. Vedremo nella prossima sezione che il valore dei rapporti è determinato in massima parte del cluster famiglia.

È ovvio che l'effetto sulla varianza totale del CSSF così come della post-stratificazione è tanto minore quanto più è grande VPS.

È interessante sintetizzare i risultati sin qui ottenuti osservando come varia l'effetto del disegno introducendo progressivamente gli elementi caratteristici dello schema di campionamento FL sulla base del dimensionamento previsto per l'Umbria. Nella tabella 4 partendo dal CCSI nelle province e dallo stimatore (7) sono riportati i valori di *deff* che si ottengono dopo aver introdotto in successione il campionamento a due stadi con la stratificazione ST1 (terza colonna), la post-stratificazione (quarta colonna), il campionamento sistematico delle famiglie (quinta colonna), la stratificazione ST2 ed ST3 (rispettivamente sesta e settima colonna).

Tabella 4 - Variazione del *deff* sotto l'azione degli elementi del disegno di campionamento FL

AGGREGATO	CCSI	PSU Comune	post-strat.	CSSF	ST2	ST3
AGR	1,00	3,67	3,65	3,94	2,12	2,03
IND	1,00	2,24	2,19	2,22	1,60	1,24
ATT	1,00	2,63	2,62	2,72	1,24	1,22
OCC	1,00	1,51	1,43	1,26	0,98	0,88
NOC	1,00	1,14	1,14	1,24	1,22	1,22
POC	1,00	1,11	1,11	1,15	1,10	1,09
DIS	1,00	1,12	1,12	1,18	1,12	1,12
FLA	1,00	1,46	1,39	1,23	0,99	0,89

La tabella evidenzia il notevole contributo positivo della post-stratificazione, del CSSF, della stratificazione ST2 e ST3 sulle variabili OCC e FLA. Purtroppo non si può dire altrettanto per le altre variabili. Quelle relative ai tre settori di attività economica beneficiano di una forte riduzione del *deff* a causa della stratificazione ST2 e, in misura minore, ST3. Circa le variabili relative alla disoccupazione la stratificazione ST2 riesce appena a compensare (solo in parte nel caso di NOC) l'effetto negativo del CSSF.

Tabella 5 - Effetto del cluster famiglia e del campionamento sistematico per diverse ampiezze del campione (media su 6 comuni)

		DIMENSIONE DEL CAMPIONE				
		25	50	100	150	200
AGGREGATO	CCSF					
AGR	1,46	0,96	0,90	0,86	0,83	0,92
IND	1,09	0,99	0,91	0,88	1,01	1,07
ATT	1,14	0,96	0,93	0,95	0,94	0,79
OCC	0,84	0,99	0,95	1,01	0,88	0,98
NOC	1,06	1,00	0,94	0,83	0,91	0,85
POC	1,05	1,00	0,96	1,02	1,10	0,83
DIS	1,06	1,01	0,94	0,98	0,94	0,75
FLA	0,89	0,99	0,98	0,98	0,87	0,94

2.2. Campionamento sistematico e cluster famiglia

Nella sezione precedente è stato valutato l'effetto del CSSF rispetto al CCSI dentro le UPSC. È interessante però tentare di separare l'azione del campionamento sistematico da quello del cluster

famiglia. Allo scopo, per ragioni di economia nei tempi di calcolo, sono stati utilizzati i dati di 6 comuni diversi per dimensione demografica (da 2.000 a 13.000 famiglie) e per attività economica prevalente (industria ed altre attività). Si supponga di voler stimare il totale comunale mediante lo stimatore post-stratificato

$$\hat{Y} = \sum_{a}^{2} \hat{R}_a X_a \qquad (8)$$

(l'universo è ora la popolazione del singolo comune). L'obiettivo formulato può essere raggiunto rapportando dapprima la varianza della (8) nell'ambito del campionamento casuale semplice delle famiglie (CCSF) con quella del CCSI misurando così l'impatto del cluster famiglia. Successivamente, calcolando il rapporto tra la varianza della (8) nell'ambito del CSSF (mediante l'estrazione di tutti i campioni sistematici di una data ampiezza) con quella del CCSF si può misurare l'impatto del campionamento sistematico.

Nella tabella 5 per ciascun aggregato è stata riportata la media dei rapporti (seconda colonna) tra la varianza del CCSF e quella del CCSI ottenuti nei 6 comuni considerati. È da aggiungere che il coefficiente di variazione dei rapporti comunali è risultato pari allo 8,95% per AGR e inferiore al 4,00% per tutti gli altri casi. È evidente che le famiglie presentano una certa tendenza alla omogeneità, soprattutto rispetto ad AGR. Al contrario, rispetto alle variabili FLA ed OCC la tendenza è verso l'eterogeneità.

Sempre nella tabella 5 sono riportati i valori medi dei rapporti tra la varianza del CSSF con quella del CCSF ottenuti nei 6 comuni per diverse dimensioni del campione. Dall'esame della tabella si evince che il campionamento sistematico risulta mediamente più efficiente rispetto a quello casuale semplice. L'efficienza sembra aumentare con la dimensione del campione anche se nel contempo essa diventa più precaria: compaiono cioè valori maggiori di uno. Infatti, per quanto riguarda la variabilità di questi rapporti si osserva un aumento della stessa al crescere della dimensione del campione. Il coefficiente di variazione infatti passa da un massimo dell'11% nei campioni di ampiezza 25 ad un massimo del 40% in quelli di ampiezza 200.

In sintesi il campionamento sistematico può essere ritenuto non peggiore del campionamento casuale semplice, anzi, il risultato sembra confermare la congettura che il campionamento sistematico dalle liste anagrafiche comporti una qualche forma di implicita stratifica-

zione delle famiglie sul territorio comunale. Un'ulteriore conclusione da trarre, tornando ai valori della quarta colonna nella tabella 3 (che esprimevano l'effetto del CSSF rispetto al CCSI) è che essi sono essenzialmente determinati dal cluster famiglia.

3. CONSIDERAZIONI FINALI

Occcorre ribadire che i risultati del presente lavoro si basano sui dati della popolazione umbra e non sono pertanto direttamente generalizzabili ad altre situazioni. L'Umbria è una delle regioni con minor popolazione e penultima in quanto a numero dei comuni. Tuttavia i risultati ottenuti forniscono indicazioni interessanti sul comportamento delle stime e suggeriscono spunti da verificare con studi similari su regioni strutturalmente diverse. Riteniamo, ad esempio, che le conclusioni tratte circa l'effetto dei vari elementi caratteristici del disegno di campionamento FL siano di carattere generale anche al di là dello specifico dato umbro. È ragionevole ad esempio pensare che la tendenza delle famiglie alla omogeneità o eterogeneità rispetto ai diversi caratteri sia una caratteristica costante sul territorio nazionale.

Il risultato concernente il rapporto fra le varianze di primo e secondo stadio è quello a nostro parere più pregnante per le implicazioni sul disegno dell'indagine: è quindi auspicabile una sua verifica soprattutto in quelle regioni dove la dimensione del campione al secondo stadio è assai più grande rispetto a quella dell'Umbria. Se poi il risultato da noi trovato risultasse confermato si impongono alcune considerazioni.

La scarsa incidenza della varianza di primo stadio su quella totale può far pensare ad un eccessivo numero di unità primarie mentre, dal punto di vista dell'efficienza delle stime, rende scarsamente rilevante la ricerca di criteri di stratificazione più sofisticati, almeno per le variabili prese in considerazione in questo studio (infatti anche dimezzando la varianza di primo stadio, la varianza totale si ridurrebbe di una percentuale pari alla metà della frazione di varianza di primo stadio moltiplicata per cento, il che darebbe qualche risultato per AGR e, in misura minore, per IND), tanto più che per l'analisi delle sottoclassi, cioè partizioni del campione, il guadagno in efficienza dovuto alla stratificazione tende ad annullarsi (Kish e Frankel, 1974).

Per quanto riguarda la stima della varianza degli stimatori del totale, il metodo del collassamento degli strati troverebbe motivo di rilancio. Infatti, se da una parte si opera come se la stratificazione fosse meno fine, dall'altra quest'ultima, andando ad agire su una frazione esigua di varianza non ha un impatto decisivo sulla varianza totale. Ne consegue che la distorsione positiva, caratteristica del metodo in questione, risulterebbe piuttosto contenuta.

Una valutazione complessiva della efficienza del piano di campionamento esaminato, potrà essere fatta solo tenendo conto dei costi della rilevazione. Per formulare un giudizio positivo o negativo occorre prima chiedersi se è possibile ridurre la varianza delle stime senza incrementare il tetto di spesa prefissato o, al contrario, diminuire la spesa senza incrementare la varianza. Ad esempio, il cluster comune e quello famiglia, pur con un impatto inflattivo sulla varianza delle stime, sono elementi di estrema semplificazione nella estrazione del campione con un conseguente minore onere. Non è così invece per la stratificazione. Se si dovesse procedere ad un ridisegno sarebbe interessante valutare l'opportunità di abbassare la soglia di abitanti oltre il quale il comune diventa autorappresentativo ed abolire il settore statistico e la zona altimetrica.

Infine i risultati evidenziano il formidabile problema rappresentato dal disegno delle indagini multiscopo quando un medesimo schema di campionamento ed uno stesso stimatore devono servire per più variabili: accade che alcune di esse traggono beneficio da elementi che invece danneggiano altre variabili, aumentandone la varianza del relativo stimatore. Inoltre per un panorama completo della situazione non va dimenticata l'importante componente dell'errore non campionario e le sue interazioni con gli elementi del disegno di campionamento.

RIFERIMENTI BIBLIOGRAFICI

W. COCHRAN, *Sampling Techniques*, John Wiley & Sons, New York, 1977.

L. FABBRIS, L. BERNARDI, *Disegno e caratteristiche dell'indagine sulle Forze di Lavoro*, Progetto FOLA, nota interna.

I.P. FELLEGI, G.B. GRAY, R. PLATEK, *The new design of the canadian Labour Force Survey*, «Jour. Am. Stat. Ass.» nr. 62, pp. 421-453, 1967.

G. B. GRAY, *Components of variance model in multistage stratified samples*, «Survey methodology», nr. 1, pp. 27-43, 1975.

Istat, *Rilevazione campionaria delle Forze di Lavoro*, «Metodi e Norme», Serie A, nr. 15, 1978.

L. KISH, *Survey Sampling*, John Wiley & Sons, New York, 1965.

L. KISK, M.R. FRANKEL, *Inference from complex samples*, «Jour. Roy. Stat. Soc.», nr. 36 (B), pp. 1-37, 1974.

R.H. HANSON, *The Current Population Survey. Design and Methodology*, «Bureau of Census, Technical Report», nr. 40, 1978.

C.A. O'MUIRCHEARTAIGH, *Response error in an attitudinal sample survey*, in «Quality and Quantity», vol. 10, 1978.

F. ZANNELLA, *Criteri per la stratificazione dei comuni nelle indagini campionarie sulla popolazione: confronti sperimentali e proposte operative*, Istat, Mimeo, Roma, 1987.

INTERVENTI

DANIELA COCCHI

Lo schema di lavoro proposto dalla commissione consente di fare qualche commento sia riguardo ad un impianto di indagine del tutto generico sia con riferimento a indagini particolari. Mi limiterò a menzionare alcuni punti che, con uno sforzo non eccessivo, potrebbero essere sviluppati.

Mi pare innanzitutto che il percorso illustrato da Zannella possa costituire una base di informazione adeguata non solo nella fase del disegno dell'indagine campionaria, ma anche nella fase successiva al campionamento. Esso appare, infatti, un ottimo punto di partenza per impostare tecniche sperimentali di stima da tentare in alternativa a quelle tradizionalmente adottate dall'Istituto. Si tratta infatti della considerazione integrata di informazioni extracampionarie, non ancora informatizzate, fino ad oggi difficili da collegare e riprodurre: ad esempio, in un lavoro che analizzava i dati della rilevazione sulle forze di lavoro in modo da tenere conto della rotazione del campione, abbiamo dovuto costruire uno schema analogo a quello appena presentato.

Inoltre, una procedura automatica come quella presentata permette di proporre con facilità stratificazioni alternative. Le prove effettuate fino ad ora hanno considerato la stratificazione attualmente usata per la rilevazione sulle forze di lavoro e l'ampiezza demografica dei comuni come variabile di stratificazione per il numero di strati fissato dalla rilevazione suddetta, in alternativa a stratificazioni equivalenti al campionamento casuale semplice entro le province.

Il livello provinciale come limite per la costruzione di strati, se ha una grande rilevanza dal punto di vista istituzionale, ha però una importanza minore dal punto di vista socio-economico, anche se fino ad ora non è mai stato valicato. Ad esempio, quando si fanno tentativi di raggruppamento di strati per stimare le varianze, e si considerano congiuntamente valori riferiti a strati diversi, di solito non si abbinano valori corrispondenti a strati, anche molto simili, che appartengono però a province diverse. Che cosa avviene, almeno a livello sperimentale, se la ripartizione provinciale non è considerata così importante e si provano anche strati che attraversano i confini provinciali?

La creazione di un file con molte caratteristiche comunali aggiornate costituisce la base per la stratificazione. Tra queste posso

no essere introdotte alcune variabili, a mio avviso molto importanti, che potrebbero rientrare nei tentativi di controllo dell'errore extracampionario, come la disponibilità della struttura amministrativa del comune allo svolgimento della rilevazione e la qualità della rilevazione nel comune stesso. Queste variabili, soprattutto la prima, condizionano la possibilità che i comuni, anche se estratti per l'indagine, di fatto entrino nella rilevazione, indipendentemente dal valore della probabilità di inclusione nella rilevazione. Altre variabili di stratificazione, su cui potrebbe valere la pena fare delle prove, sono le variabili indicatrici della suddivisione dei centri metropolitani e di diverse tipologie familiari.

Sulla questione del numero di interviste per rilevatore, se si passa ad indagini mensili e si toglie la settimana di riferimento, si avrebbe una conseguenza importante: prestazioni di tipo occasionale e precario potrebbero trasformarsi nella professione di rilevatore.

La relazione di Fabbris potrebbe, a sua volta, rivelarsi un ottimo strumento per affrontare la questione del sovracampionamento. Nell'indagine sulle forze di lavoro, il sovracampionamento, quando c'è stato, ha avuto l'effetto di raddoppiare il numero dei comuni per strato, estraendo poi le famiglie al loro interno senza alterare la frazione di campionamento. Qual'è, con la innovazione metodologica proposta da Fabbris, la riduzione nella variabilità delle PSU dovuta all'estrazione di due comuni? In che misura può avere senso aumentare il numero di comuni per strato e diminuire il numero di famiglie? Poiché ci sono numerosi segnali di insoddisfazione nei riguardi del sovracampionamento nella sua configurazione attuale, quale tendenza si può suggerire, aumentare il numero di comuni per strato o aumentare il tasso di campionamento in certe zone? Mi sembra di poter dire che il lavoro di Fabbris, con qualche integrazione, permetterebbe di analizzare fino a che punto si possa sfruttare la numerosità per strato pari a 2 per migliorare l'efficienza delle stime e valutare che cosa abbia dato il sovracampionamento a questo riguardo. Naturalmente si dovrà tenere conto dell'esperienza di Montanari, che indaga invece su quale sia l'inerzia delle variabili sulle due fonti di variazione, interrogandosi se valga veramente la pena sovracampionare, soprattutto per variabili come la disoccupazoine, la cui varianza è essenzialmente di secondo stadio.

Un'ulteriore considerazione riguarda la varietà degli ambiti territoriali considerati nelle diverse sperimentazioni a me note (Piemon-

te-Toscana-Calabria, Lombardia-Veneto, Umbria, Emilia-Romagna), che non consente confronti diretti tra metodi e strategie. La sperimentazione delle diverse proposte dovrebbe essere effettuata sullo stesso campione.

Per concludere, la mia sensazione sul problema del miglioramento dell'efficienza delle stime ottenuto per mezzo della stratificazione è che, per risposte a livello locale, per molte variabili, piuttosto che verso una stratificazione ottima, le soluzioni interessanti saranno orientate verso l'impiego di metodi per piccole aree.

GIUSEPPE CICCHITELLI

Come membro della Commissione campioni, sono in una certa misura corresponsabile dell'attività svolta e delle elaborazioni presentate oggi.

Interverrò su qualche aspetto dei due documenti illustrati, partendo dall'ultimo.

Il Prof. Fabbris esplora le numerose tecniche di formazione del campione con probabilità variabili e proporzionali alle dimensioni, che sono state proposte in letteratura.

L'obiettivo è quello di selezionare un gruppo ristretto di procedure, adottabili nelle indagini campionarie Istat. I parametri di riferimento, per un giudizio, sono sostanzialmente tre: la stabilità delle stime, che si può chiamare efficienza, la stabilità dello stimatore della varianza delle stime e la praticabilità, anche in termini informatici, delle procedure di formazione del campione e di stima. Dalle cinquanta tecniche, prese in esame, sulla base di un primo giudizio di adeguatezza, si arriva a selezionarne dieci, che vengono confrontate, mediante un'ampia sperimentazione su dati reali di tre regioni e con riferimento ad alcune significative variabili.

Devo dire che il lavoro è stato arduo e assai impegnativo anche dal punto di vista della realizzazione delle relative procedure informatiche.

Ritengo che i risultati siano molto interessanti e che trascendano l'immediata operatività in campo Istat, nel senso che costituiscono un contributo anche metodologico.

La motivazione che è alla base della ricerca è la tesi secondo cui, nelle indagini campionarie sulla popolazione, è preferibile sele-

zionare due comuni o più per strato, anziché uno, come avviene nell'indagine trimestrale sulle forze di lavoro e in molte altre indagini occasionali Istat.

Militano a favore di questa soluzione due argomenti. In primo luogo, la possibilità di stimare correttamente la varianza di primo stadio, senza ricorrere a procedure approssimate come il collassamento degli strati.

Secondo, la possibilità di misurare l'errore dell'intervistatore mediante la tecnica dei campioni compenetranti, tecnica che non è compatibile con la selezione di una unità per strato.

Fra queste due motivazioni, darei la preminenza alla seconda, perché la prima questione, cioè la possibilità di stimare correttamente la varianza degli stimatori, credo che non sia di grossissima rilevanza, visto che la tecnica del collassamento degli strati fornisce una ragionevole soluzione approssimata.

A favore della selezione di un solo comune per strato c'è invece l'argomento della semplicità del procedimento e dell'opportunità di stratificare più finemente la popolazione.

Certo, le ragioni a sostegno dell'opzione dei due comuni per strato possono non apparire decisive e nette, e, probabilmente, un confronto di efficienza con l'impostazione attuale avrebbe potuto dare qualche illuminazione in più. Del resto, la problematicità della scelta è anche testimoniata nella prassi, dal fatto che i due Istituti di Statistica più autorevoli a livello mondiale, quello statunitense e quello canadese, adottano strategie opposte. Una unità primaria per strato il primo, due o più il secondo.

Tuttavia, nell'ottica del disegno generalizzato di campionamento nelle indagini Istat, è sicuramente importante poter disporre di precise indicazioni circa le procedure statistiche e informatiche da adottare ogni volta che si volesse innovare circa il numero delle unità primarie da selezionare per ogni strato. La scelta, poi, della tecnica specifica, tra le poche superstiti al setaccio del Prof. Fabbris, potrebbe essere basata più che su una comparazione stretta in termini di efficienza, su una questione di praticabilità. Ciò in ragione del fatto che da analisi empiriche risulta che la variabilità di primo stadio rappresenta una quota modesta della variabilità complessiva delle stime almeno per alcune variabili rilevanti. Al riguardo, il Dottor Montanari riferirà sui risultati di uno studio condotto per il campione forze di lavoro dell'Umbria con i dati dell'ultimo censimento.

D'altra parte, va ricordato che il campionamento con probabilità variabili non ha solo come finalità quella di ottimizzare le stime; esso può essere finalizzato a far sì che in un contesto di autoponderazione, in presenza di strati di ampiezza pressoché uguale, il numero di unità finali da intervistare per ogni unità sia costante. Cosa che può essere rilevante per una razionale politica di assegnazione delle interviste ai rilevatori.

Farò una ultima notazione su un punto specifico: ho qualche perplessità sulla definizione di stimatore per quoziente, scritto come rapporto tra somma campionaria della variabile oggetto di studio e somma delle probabilità assegnate alle unità nell'ordine in cui vengono selezionate. Credo che al denominatore vadano le probabilità iniziali a prescindere dall'ordine di estrazione; questo fatto, potrebbe incidere in modo sistematico, in qualche misura, sul comportamento dello stimatore. Tale aspetto è forse marginale, anche perché lo stimatore per quoziente, così come è costruito, usa l'informazione ausiliaria (dimensione demografica) due volte: una prima volta a livello di formazione del campione ed una seconda volta a livello di stima. Quindi, tutto sommato, i giudizi sui confronti tra le varie procedure non dovrebbero cambiare.

Per quanto riguarda il documento di Zannella, devo dire che esso consente di valutare *ex ante* l'efficacia di variabili di stratificazione, nonché di procedimenti di raggruppamento alternativi, fondati su informazioni a livello di comune, dedotte dal censimento, dati talora aggiornati, come la popolazione residente.

Ho manifestato, nel dibattito in Commissione, che si tratta di un approccio molto interessante: la stratificazione è ancorata ad elementi obiettivi ed è disegnabile *ad hoc*, secondo la tipologia della specifica indagine.

Il lavoro di sperimentazione è stato condotto, utilizzando una formula di varianza semplificata rispetto a quella che richiederebbe la struttura complessa del campione Istat. Cioè, in sostanza, si assume il campionamento casuale semplice delle unità elementari, anziché quello sistematico delle famiglie, come avviene nella realtà. Questa strada che, peraltro, non ha alternative praticabili, va considerata comunque valida per l'ottenimento di indicazioni generali, almeno per diverse delle variabili considerate. Ciò, anche alla luce di alcuni risultati ottenuti per la regione Umbria che pure saranno illustrati dal Dottor Montanari. Fra gli altri risultati che emergono dalle

elaborazioni numeriche, c'è la pratica irrilevanza del settore statistico come base territoriale per la stratificazione. Ciò è un ulteriore elemento che milita a favore della cancellazione di questa entità territoriale, che ha un sapore un po' archeologico, anche per l'indagine sulle forze di lavoro. La propensione di Zannella verso la stratificazione mediante classi di popolazione, con il vincolo di un uguale ammontare di popolazione negli strati è supportata da alcune risultanze della sperimentazione, ma soprattutto da considerazioni sulla praticabilità. Questo punto di vista, mi pare senz'altro condivisibile.

ALFONSO ORSI

Vorrei ringraziare il vertice dell'Istat e il professor Leti per averci invitati a questa riunione di notevole interesse. Tale interesse sollecita anche un approfondimento dal punto di vista settoriale; sarebbe importante riprendere questa discussione e questa analisi metodologica e applicativa, non soltanto cioè dal punto di vista così generale, ma anche settoriale, distintamente demografico, sociale, ed economico. Ricordo che negli anni più giovanili erano molto frequenti questi incontri, proprio per discutere i problemi dei campioni da un punto di vista applicativo, dal punto di vista di ciò che effettivamente facevamo e traendone quindi le conclusioni operative.

Intanto, l'interesse dei campioni per l'agricoltura, in modo particolare. Credo sia noto a tutti l'importanza del giorno in cui il Dipartimento dell'Agricoltura degli Stati Uniti di America «tira fuori» il dato sulla produzione del frumento risultante dalle rilevazioni campionarie. E come questo dato sia tenuto gelosamente custodito fino a quel momento e quale influenza esso abbia sulla Borsa, quali ripercussioni, quali implicazioni commerciali abbia sui prezzi, etc. Ricordo che il nostro era (anni cinquanta, inizio sessanta) uno di quei due Paesi che più avevano sviluppato il sistema campionario. Proprio nel campo demografico ed in quello dell'agricoltura avvennero in Italia le prime applicazioni campionarie: indagini sulle famiglie e sulle coltivazioni attraverso «l'area sampling», il Capo Servizio delle Statistiche Agricole era, pro-tempore, anche Capo dell'Ufficio studi.

Per il Servizio Statistiche Agricole c'era una notevole assistenza da parte di funzionari, matematici esperti del campione dell'Ufficio studi. Ma c'era anche un distacco organico di funzionari: Guarini, Esposito, Bartoli e non a Livello di Servizio, ma anche a livello di Reparto.

Si discutevano i risultati del campione, si arrivava a miglioramenti che completavano l'opera meritoria dell'Ufficio studi, nel quale operavano altri miei giovani colleghi. Ecco: le applicazioni della tecnica campionaria ci consentiva, distintamente il 15 maggio, il 15 giugno e il 10 di ottobre, di far pervenire un telegramma al Ministro dell'Agricoltura; per le prime due date, i dati di previsione, fatti con l'area sampling, di accertamento del frumento e, aggiungo, successivamente anche di altre coltivazioni.

La parola «diaspora» che ha usato il professor Leti, stamattina, quando parlava appunto di diaspora di funzionari dal Reparto studi, degli specialisti cioè del progetto campioni, va un pochino corretta.

Infatti io parlerei invece non di diaspora bensì di «inseminazione» che spero sia feconda e che interessi sempre più altri Servizi. Nulla toglie che si debba poi ricostituire anche quella unità all'interno del Reparto studi, ma il suo distacco non lo si deve vedere solo come una diaspora! Quelli che diasporano, poi, creano notevoli benefici agli altri Servizi!

In merito alla relazione di Zannella (il quale ha richiamato il fatto che talune caratteristiche nell'unità di rilevazione, se numerose, sono difficili a cogliersi) vorrei brevemente intervenire.

Questo, infatti, si verifica soprattutto nel caso della azienda agricola ed è vero. Normalmente campioni agricoli sono basati su di una stratificazione fatta (per l'indagine sulla struttura delle aziende agricole) in termini di superficie agricola utilizzata. Sulla base di tale stratificazione, pretendiamo di cogliere oltre le principali coltivazioni tanti altri aspetti, che si colgono bene col campione, dal punto di vista qualitativo, ma da un punto di vista quantitativo lasciano qualche perplessità. Probabilmente noi siamo troppo attratti dal fascino dei campioni unici, polivalenti, etc. Nel corso della mia esperienza in campo internazionale mi è capitato di vedere che altrove danno meno importanza a campioni polivalenti, ma fanno più «linkage» di campioni diversi con diverse finalità, che però, hanno una matrice di stratificazione comune anche. Così, ad esempio, la superficie agricola utilizzata viene combinata per altri campioni con ulteriori stratificazioni. Questo avviene soprattutto nei Paesi del Centro e Nord Europa. Si potrebbe dire: ma lì hanno poche colture, hanno pochi allevamenti e tutto è molto più semplice. Però, proprio questo motivo potrebbe spingerci a studiare quest'altra strada.

Un altro punto su cui desideravo richiamare l'attenzione dei colleghi soprattutto i metodologi, (ai quali va il ringraziamento mio e del Servizio) è il fatto che il rapporto dovrebbe essere ancora maggiormente appropriato tra organizzazione periferica per la raccolta dei dati e la sua influenza sulla metodologia e sulla tecnica del campione.

Ci sono dei rapporti reciproci e, ovviamente, bisognerà studiare attentamente di far sì che l'organizzazione periferica della raccolta dei dati abbia condizionamento minimo, cioè che condizioni il meno possibile nella determinazione della metodologia.

A mio avviso, dovrebbe essere la metodologia che condiziona la organizzazione periferica e non viceversa.

Anche questo è uno dei problemi connessi con i campioni. Mi scuso se sono stato molto breve, se ho portato a questo incontro la mia esperienza quasi quarantennale: ho cominciato ad occuparmi di campioni da quando ero studente, poi come assistente all'Università. È un contributo che ritengo utile. Grazie.

ROLANDO ANGELONI

La Commissione per lo studio dei campioni ha svolto una notevole mole di lavoro, però, a mio avviso, non ha forse considerato nel suo giusto valore il rapporto tra studi teorici e realtà pratica.

Per esempio, un campione a uno stadio per la rilevazione di aziende con bestiame suino o bovino è sempre disperso in un gran numero di comuni e tra questi ve ne sono alcuni, talora non pochi, con non più di 2-3 unità da rilevare. Poiché in pratica il comune non trova un rilevatore che si impegni per sole 2-3 aziende (data l'entità dei compensi) ci si trova di fronte alle seguenti alternative:

1) Ricorrere a rilevatori (eventualmente professionisti) che vadano in giro per i diversi comuni in modo da rilevare un numero di aziende da concretare una convenienza economica: è quello che accade nelle rilevazioni relative al prodotto lordo delle imprese dei rami 2, 3, 4, 5, 6, 7, 8 e aventi da 1 a 9 addetti;

2) Rinunciare alla rilevazione in tutti i comuni con 2-3 aziende;

3) Fare un campione a due stadi.

La prima soluzione comporta un aumento dei costi, la seconda una perdita di informazioni, la terza un aumento — a parità di numero di aziende — dell'errore campionario.

In pratica è stata adottata la terza soluzione, ma il problema non si porrebbe se si disponesse di una rete di rilevatori. Mi sembra allora che la Commissione potrebbe esaminare l'opportunità, per l'Istat, di creare una propria rete di rilevatori, analizzando i costi ed i benefici per vedere se ciò conviene, tenendo in particolare presente quanto si spende attualmente per le varie indagini su campo e quanto se ne ricava in termini di modelli validi.

Altro esempio: per l'indagine sul prodotto lordo 1986 delle imprese con 1-9 addetti il campione è stato preparato su base censimento 1981.

In sede di rilevazione, avvenuta a cavallo degli anni 1987-88, la base ha rilevato un grosso logorio: le imprese che avevano cambiato indirizzo o attività o classe di addetti o avevano cessato l'attività non erano poche, e malgrado si sia fatto ricorso largamente agli elenchi suppletivi, su un campione di circa 30.000 imprese siamo riusciti a portare a casa solo 21-22.000 modelli validi.

Tralasciando il problema dell'aggiornamento degli schedari in genere, quel qui importa è che, per l'indagine sul prodotto lordo 1988 si dovrà utilizzare lo stesso campione relativo al 1986 integrato da altre unità estratte ancora dalla base censuaria dopo averla depurata dei campioni base e suppletivo estratti per il 1986.

Questa depurazione comporta una variazione dell'universo da cui si estrae il campione integrativo, e quindi le unità di tale campione vengono estratte con probabilità diversa da quelle del campione 1986 estratto dall'universo totale. Ciò dà origine a qualche problema per il riporto all'universo, problema che converrà studiare se si vogliono evitare distorsioni nelle stime. Si tratta di un tipico problema teorico originato da inconvenienti pratici, ed è anche questo tipo di problemi che la Commissione dovrebbe studiare.

AMATO HERZEL

Sono uno dei membri della Commissione, ma non per questo eccessivamente condizionato dai documenti che sono stati presentati. In tali documenti c'è l'impronta degli autori che prevale sull'appartenenza alla Commissione. Quest'ultima infatti ha stimolato e approvato il progetto di vari lavori ma poi, mi pare giustamente, non ha preso posizione ufficialmente su tutti i contenuti dei documenti. Questa precisazione mi sembra opportuna per evitare fraintendimenti circa la scelta della presente.

Ognuno ha lavorato in piena autonomia, sia pure nell'ambito di sforzi speriamo convergenti, ma ovviamente i lavori esprimono anche posizioni diverse, come si è potuto constatare dagli interventi precedenti in cui sono emerse opinioni diverse su vari problemi, fra cui mi sembra importante quello riguardante il numero delle unità primarie da estrarre dagli strati non autorappresentativi.

Sarà magari per una deformazione professionale, ma per me la questione della stima della varianza di uno stimatore è importante, anche perché ha tanti riflessi pratici evidenti. È chiaramente legata alla questione della determinazione della dimensione del campione, è legata alla scelta di vari procedimenti sia di stima, sia di piani di campionamento.

Se estraiamo una unità da uno strato, non solo non esiste uno stimatore corretto della varianza, ma in realtà a volere essere rigorosi non esiste nessuno stimatore, perché il procedimento degli strati collassati ha una sua parziale giustificazione nei casi in cui si estraggono negli strati le unità con probabilità uguale, il che non avviene nei nostri schemi.

Non credo che la questione possa essere considerata trascurabile, anche perché un errore sistematico nello stimatore della varianza della stima si ripercuote anche sul calcolo dei «deff», in quanto incide sulla varianza dello stimatore del campionamento semplice corrispondente. D'altra parte non ci sono, in realtà, dei motivi veramente rilevanti per scartare l'ipotesi di estrarre almeno, dico almeno, due comuni su tutti gli strati non autorappresentativi.

La preferenza che è stata data in passato al procedimento di un campione di una unità per strato ha probabilmente soprattutto delle motivazioni storiche; fino a non molto tempo fa, non si conoscevano procedimenti efficienti per estrarre, senza reimmissione, due o più unità con probabilità variabili; adesso invece ce ne sono diversi, come vediamo del resto dal rapporto Fabbris, quindi questo ostacolo ora viene a cadere.

Per quanto riguarda il timore di perdere efficienza riducendo il numero degli strati, gli stessi risultati di Fabbris, come anche le elaborazioni di Zannella, dimostrano che questo rischio è praticamente inesistente. È vero, come dice Russo, che la varianza non la possiamo stimare ugualmente in modo corretto neanche con una o più unità per strato, perché gli stimatori che impieghiamo sono degli stimatori post-stratificati e stimatori quozienti; esistono, comunque, in questo caso, degli stimatori approssimati soddisfacenti.

Comunque, non si vede, a mio parere, perché anche ammesso che ci sia una distorsione nella stima della varianza dovuta alla natura dello stimatore delle medie, o dei totali, si debba aggiungere a questo un altro elemento che introduce un'altra distorsione che può benissimo cumularsi con la prima.

Inoltre, sono convinto che sarebbe estremamente utile effettuare delle compenetrazioni al fine di poter migliorare la qualità dei dati. Anche per questo motivo sono fautore, in linea di massima, di un campionamento di almeno due unità primarie per ogni strato e devo dire che, sotto questo profilo, il lavoro che ha svolto Fabbris, con i suoi collaboratori o ex collaboratori dell'Istat, è di altissimo valore.

È veramente una massa enorme di dati, e sono dati utili e importanti che meriterebbero, anzi richiederebbero, senz'altro una diffusione non limitativa solo all'interno dell'ambiente dell'Istat ma più larga, perché in tutta la letteratura, a quello che mi risulta, non esistono studi di queste dimensioni. Circa le proprietà degli stimatori che si ottengono con i campionamenti con probabilità variabili, Fabbris ha anche dimostrato in questo modo che l'Istat può assolvere ad una funzione importante, dal punto di vista della verifica, diciamo pure sperimentale di vari disegni e procedimenti che vengono proposti, perché l'Istituto dispone oltre a tutto di una ricchezza di dati da permettere confronti ed elaborazioni che in altre sedi non sarebbero possibili.

In questo senso, penso che Fabbris indichi una strada ed è auspicabile che l'Istat la voglia in futuro proseguire. D'altra parte lo stesso Fabbris ci dice che non ritiene terminato il suo lavoro ed anzi dà nell'ultima parte dei suggerimenti per continuare.

Vorrei inserirmi in questi suggerimenti segnalando che, anche se dieci procedimenti di estrazione con probabilità variabili trattati da Fabbris possono sembrare molti, ne sono stati trascurati altri che sembrano molto importanti. Per esempio mi riferisco al procedimento di Singh per il campione di due elementi e a quello di Sampford per il campione di tre unità. Dal lavoro di Fabbris andrebbe messo in evidenza un punto importante: dalle sue elaborazioni risulta che in un campione sistematico, però non casuale, la varianza della varianza possa essere stimata come nel campionamento sistematico casualizzato.

Sarebbe un risultato di grande interesse, ma dovrebbe forse essere verificato ulteriormente.

Un altro punto, ancora, mi permetto di suggerire: converrebbe pensare ad altri tipi di variabili. Le ricerche di Fabbris sono state condotte su delle frequenze assolute che sono certamente correlate con le dimensioni delle unità campionarie. Sarebbe interessante vedere che cosa succede quando si prendono altri caratteri che siano meno correlati o addirittura negativamente correlati. È facile prevedere che le cose andranno meno bene, forse addirittura andranno molto peggio, comunque si avranno dei risultati di tipo diverso. Basterebbe forse anche soltanto prendere al posto delle frequenze assolute delle frequenze relative (ad esempio, le percentuali di disoccupati) per vedere se e in che misura i risultati cambiano. Sono stato tirato in ballo anche a titolo personale; Fabbris ha voluto gentilmente riprendere una proposta piuttosto marginale contenuta in un mio articolo e ha considerato i risultati che si ottengono applicando una certa formula. Questi risultati sono talvolta buoni e talvolta cattivi; a me nel complesso sembrano abbastanza soddisfacenti e i giudizi che ne dà Fabbris non sono univoci. Certamente non credo che sia un procedimento consigliabile in tutti i casi, perché presenta l'inconveniente di dare luogo, qualche volta, a probabilità negative. Queste possono essere eliminate, ma con un lavoro supplementare che certamente male si concilia coi criteri di semplicità e praticità che giustamente Fabbris pone a base della scelta del piano di campionamento. D'altra parte credo che le probabilità negative si manifestino soprattutto quando gli strati contengono unità di dimensioni molto diverse. Quindi questa evenienza potrebbe anche essere usata come un indice, che ci dice che c'è qualcosa da rivedere in quei tipi di strati: che forse sarebbe il caso di dividerli per renderli più omogenei.

L'argomento principale che porta Fabbris circa la non applicabilità delle mie formule riguarda la memoria che viene richiesta e qui mi pare che ci sia un equivoco: queste formule sono espresse, come si usa dire, in termini di campioni considerati come insiemi, non come successioni, ma le formule quando funzionano, quando cioè non ci sono probabilità negative, possono essere molto facilmente trasformate come è ben noto, dalla teoria generale del campionamento con probabilità variabili, in procedimenti «draw by draw», in cui si estraggono le unità una per volta, il che riduce enormemente l'impiego della memoria del calcolatore. Vorrei infine soffermarmi brevemente sulla questione se convenga estrarre 2 o 3 unità per strato. È un argomento che andrebbe considerato con maggiore atten-

zione; se si considerano i dati prodotti da Fabbris, si vede che c'è un netto miglioramento per quanto riguarda la stabilità dello stimatore del totale o dello stimatore della sua varianza. Su questo punto in letteratura c'è un equivoco abbastanza diffuso; c'è un lavoro abbastanza noto in cui viene dimostrato che in certi casi, quando si passa da un campione di due elementi a un campione di tre elementi, le cose peggiorano in certi schemi di campionamento senza ripetizioni con probabilità variabili. Si tratta appunto di un equivoco, in quanto, nei casi considerati in quel lavoro, solo formalmente il passaggio è da due a tre elementi; in realtà cambia lo schema di estrazione.

Quando lo schema di estrazione è effettivamente uguale, nella sua struttura, allora si hanno gli stessi vantaggi in termini di stabilità che si osservano nel campionamento casuale semplice.

GIORGIO MARBACH

La Commissione ha concentrato i lavori su un tema particolare, costituito dalla problematica della scelta delle unità di primo stadio. Ha naturalmente fatto un lavoro egregio, forse il più approfondito mai effettuato su tale aspetto.

Ma la problematica dei campioni esige che ci si misuri su una serie di altri problemi. Per esempio, la Commissione si è occupata della scelta di unità di primo stadio all'interno della stratificazione, considerata come dato e non a sua volta da verificare.

Inoltre, in tema di campionamento esistono numerosi altri temi rilevanti: auspico che la Commissione continui a lavorare, per diventare anche un pò il cuore di alcune scelte strategiche dell'Istituto Nazionale di Statitisca, sulle quali mi permetto di avere un dubbio di fondo, che proprio questa riunione accentua.

La Commissione potrebbe soffermarsi in modo forte e fermo sui problemi della validità delle principali ricerche campionarie dell'Istituto: quella sulle forze di lavoro, l'indagine sui consumi delle famiglie, le indagini sui problemi delle imprese, etc. che hanno urgentemente bisogno di qualcosa di più forte di un semplice maquillage.

La variabile di base, ossia la popolazione residente o le famiglie, richiede attenzione critica del tutto particolare. I risultati che lo stesso Istat fornisce sulla base dell'indagine multiscopo delle fa

miglie, in cui si paragonano i nuclei di fatto con quelli residenti, indicano una differenza di un milione, pari a quasi tre milioni di italiani. Se aggiungiamo una stima di due milioni di stranieri abbiamo, grosso modo, cinque milioni di persone, cioè un 10% della popolazione italiana complessiva, sul quale le indagini, così come vengono fatte, non possono misurarsi, perché il punto di riferimento della stratificazione, appunto, è la popolazione residente.

Questo problema si aggraverà in futuro, se, come sembra, il prossimo censimento non sarà più occasione per una «pulitura» delle anagrafi. *I nostri campioni daranno, nella migliore delle ipotesi, una buona fotografia di un'urna distorta.*

Anche l'urna delle imprese, per quello che oggi ne sappiamo, è abbastanza distorta.

Mi soffermo ora, in particolare, sui criteri di stratificazione della indagine sulle forze di lavoro. Essi appartengono alla vetero-cultura dell'Istat.

Settore statistico, zone altimetriche, attività economica prevalente, sono aspetti che, in termini di correlazione con le principali variabili della occupazione, lasciano veramente molto perplessi. Inoltre i dati sono del 1981!

Occorre una rifondazione della indagine sulle forze di lavoro, per poter attribuire la necessaria validità e credibilità ai risultati. La scelta del collettivo di riferimento e quella dei criteri di stratificazione mi sembra abbiano carattere di priorità. È indispensabile individuare caratteri di stratificazione fortemente correlati con l'oggetto della ricerca, e continuamente aggiornabili, così come si insegna agli studenti.

In generale, ci renderemo sempre più conto che per ottenere buone indagini campionarie occorrono buone statistiche di partenza, riferite alla giusta dimensione territoriale. L'Istat, invece, attribuisce un basso grado di priorità alle elaborazioni di statistiche disaggregate territorialmente.

Confido che presto ci sia un inizio di ravvedimento. Trovo spunti interessanti nella sintetica relazione di G. Leti sul tema dei campioni areali. Il collega scrive che non è stato possibile puntare su tale impostazione, perché le sezioni di censimento sono state formate in base a criteri diversi a quelli necessari per una buona base statistica. Ma chi, se non l'Istat, ha capacità, tecniche ed il dovere

di controllare le caratteristiche della minima entità territoriale del censimento, in una impostazione strategica che è quella delle successive indagini per campione?

La Commissione ha avuto, probabilmente, un ottimo effetto nel riproporre l'avvio di una cultura del campionamento nell'Istat, che da sempre si orienta sui grandi campioni.

Un altro aspetto. L'occasione del prossimo censimento potrebbe essere molto importante per studiare in modo approfondito la differenza tipologica delle persone e delle famiglie nelle aree in cui il telefono esiste ed in quelle prive di telefono, per consentire una corretta interpretazione dei risultati delle indagini telefoniche.

Infine, un immane problema per la Commissione, una vera e propria sfida per i prossimi anni, è la stima dell'errore complessivo, campionario e non. Poiché la scelta della tecnica di campionamento non è più indifferente al problema dell'errore complessivo, occorre attentamente valutare, per esempio, se la indagine sulle imprese debba ancora essere effettuata per posta.

VINCENZO SIESTO

Marbach suggerisce, in fondo, ciò che noi vorremmo attuare da alcuni anni a questa parte, ossia delle indagini basate sull'impiego di piccoli campioni. Allora rivolgo a Marbach la domanda: la selezione di due comuni per ogni strato è compatibile con disegni di piccoli campioni?

GIORGIO MARBACH

Occorre evitare di pensare alla unità territoriale costituita dal comune. Ritengo, invece, che l'Istat abbia la possibilità di crearsi una maglia territoriale di riferimento che prescinda dall'unità amministrativa. L'Istituto può concentrare tutte le informazioni di censimento ed altre ancora, aggiornandole continuativamente.

A questo punto si può costituire rapidamente un piccolo campione, su un'area ridotta; si potrebbe ridurre notevolmente il problema delle mancate risposte ed ottenere in tempi molto rapidi risultati su qualsiasi tema.

L'Istat ha l'opportunità di liberarsi delle pastoie rappresentate dalle unità amministrative, tipo i comuni, che non hanno significato economico né sociale; molti comuni italiani hanno le dimensioni di due palazzine. L'unità provinciale non rappresenta un'area omogenea univoca e le stesse regioni non presentano omogenei profili socio-economici al proprio interno.

In conclusione, per ottenere buoni campioni, rapidamente, di ridotte dimensioni, probabilmente è necessario uno sforzo culturale, per superare l'attuale ripartizione territoriale, con salvaguardia del raccordo a comuni, province e regioni. Ma occorre mettersi su questa strada, perché avremo crescenti difficoltà per costituire le urne giuste da cui estrarre i campioni.

Per quanto riguarda l'unità impresa, l'anagrafe di quelle degne di questo nome può essere costituita attraverso sforzi congiunti di tutti coloro che si occupano di questo problema (Unioncamere, Istat, Ministero delle Finanze, Cerved, etc.) per gestire un'urna affidabile di imprese ed unità locali.

Dobbiamo puntare sul territorio, su unità territoriali piccole, raccordarle agli attuali confini amministrativi, operare attraverso una serie di panel, ossia di campioni continuativi, per poter rispondere rapidissimamente a qualsiasi esigenza, usando strumenti idonei di rilevazione ed input, e dismettendo reti di intervistatori sui quali l'Istat non abbia pieno e diretto controllo.

IPPOLITO SANETTI

I colleghi metodologi hanno fatto un lavoro encomiabile per presentarci nuove proposte di tecnica campionaria che però, a mio avviso, non sono veramente innovative in quanto nella letteratura sulla teoria dei campioni rappresentano semplicemente un perfezionamento e un affinamento sia pure apprezzabili delle vecchie tecniche fondate sulla stadificazione del campione e sulla stratificazione dell'universo.

A questo proposito e con riferimento al campione dell'indagine delle forze di lavoro — che conosco molto bene — posso citare due esperienze fatte in passato da studiosi della materia (vedi, M. Zani per la Emilia-Romagna e L. Biggeri per la Toscana) che hanno applicato la cluster analysis (utilizzando un discreto numero di variabili)

ai criteri di stratificazione dei Comuni in sostituzione dei pochi parametri del campionamento Istat (provincia, dimensione demografica, zona altimetrica e attività economica prevalente). Ebbene, entrambi sono pervenuti alla conclusione che in fin dei conti le loro stime non erano più efficienti o comunque migliori di quelle corrispondenti costruite dall'Istat.

Un elemento veramente innovativo nella costruzione dei campioni potrebbe essere quello di trovare — mi riferisco sempre all'indagine delle forze di lavoro — un metodo per riuscire a campionare quella parte di popolazione (c'è chi dice che raggiunge attualmente i 2 milioni di unità) che resta automaticamente esclusa dal campo di osservazione tutte le volte che ci si rifà alle anagrafi comunali come urne da cui estrarre i campioni, cioè quella dei non residenti.

Il campionamento areale potrebbe essere un metodo molto attuale per risolvere il problema. D'altra parte, sappiamo benissimo che non è facile costruire campioni areali: esperienze in tal senso condotte dall'Istat stesso (peraltro limitatamente alle «unità economiche» e non alle «famiglie», che a mio parere potrebbero dare esito diverso) si sono rivelate sostanzialmente negative, per diversi motivi, non ultimo quello della eccessiva dimensione attuale delle «sezioni di censimento» e della conseguente eccessiva variabilità delle unità da campionare in ciascuna sezione. Ma a questo specifico aspetto negativo si può cominciare a ovviare con la riduzione delle dimensioni delle sezioni citate, cosa che peraltro, come è noto, è già allo studio da parte dell'Istat.

In conclusione, bisogna che le nuove tecniche campionarie siano orientate a risolvere piuttosto questi ultimi problemi, affinché una qualsiasi indagine per campione raggiunga il suo scopo, che è quello di stimare con sufficiente fedeltà l'universo oggetto di osservazione senza rivelare grosse carenze informative su di esso e senza distorcerne la reale consistenza e struttura.

GIUSEPPE LETI

Desidero fare alcune precisazioni che scaturiscono dagli interventi di chi mi ha preceduto.

Come ho detto nell'introduzione ai lavori, la Commissione ha approfondito i problemi delle indagine campionarie di media e gran-

de taglia e si propone di trattare in futuro le problematiche dei piccoli campioni. È ovvio che i risultati a cui si è pervenuti non possono essere attribuiti a tipi di campionamento che la Commissione non ha ancora studiato.

La Commissione, non ha trascurato il campionamento areale, di cui anzi, come si è detto nell'introduzoine, ha iniziato a studiare vari problemi che in un secondo momento è stata costretta però, per vari motivi, ad accantonare. E ciò anche, perché la Commissione ha concentrato la sua attenzione sui campioni da liste, essendo per lo più in Italia disponibili liste su cui effettuare il campionamento. È ovvio che quando si hanno liste passa in secondo piano il campionamento areale. Lo studio di L. Fabbris è valido per tutti i campioni a due stadi in cui per ciascuna unità del primo stadio sono disponibili liste delle unità elementari. I risultati dello studio sussistono quindi anche se l'unità del primo stadio è il comune, ma essi hanno una validità del tutto generale potendo essere le unità del.primo stadio di qualunque tipo.

Un ringraziamento particolare debbo rivolgere al prof. Siesto perché ritengo che il migliore apprezzamento dei lavori della Commissione sia dovuto a lui. Infatti egli ha valutato «controcorrente» il nostro operato diverso cioè da quanto è stato fatto finora nell'Istat e diverso da ciò che si fa negli Istituti esteri di statistica. Poiché il lavoro della Commissione è stato sempre aderente alla realtà, dalle parole di Siesto rilevo la constatazione che la Commissione non ha rappezzato cose vecchie, ma è voluta uscire dalla accidiosa routine per porre le basi di nuove procedure che consentano il ricorso, più frequentemente e più appropriato, alle indagini campionarie.

FRANCESCO ZANNELLA

Nello svilupare la procedura generalizzata si è dovuto tener conto dei vincoli amministrativi ed organizzativi, che hanno fortemente condizionato le scelte metodologiche. In particolare l'elevata variabilità dell'ampiezza demografica dei comuni italiani, con una popolazione residente che va da poche centinaia a milioni di unità, pone dei notevoli problemi nella programmazione del disegno campionario.

Al termine della relazione che ho presentato, nello svolgere le considerazioni conclusive, formulo la proposta di poter superare que-

sto tipo di organizzazione mediante la creazione di pseudo unità di primo stadio. Queste pseudo PSU dovrebbero essere ottenute mediante opportuni raggruppamenti dei comuni di piccole dimensioni e suddivisioni dei comuni più grandi, in modo da ottenere unità di primo stadio della stessa ampiezza demografica. Ciò consentirebbe di risolvere numerosi problemi e di pervenire a disegni campionari molti più efficaci degli attuali.

Una seconda considerazione che intendo svolgere, riguarda le dimensioni dei campioni che vengono utilizzati dall'Istat.

Da più parti vengono mosse critiche che si muovono in direzioni diverse: «la numerosità del campione è troppo elevata» e «la numerosità non è sufficiente».

Credo che l'aspetto della dimensione campionaria debba essere affrontato in stretto collegamento con gli obiettivi dell'indagine per i quali il campione è stato programmato. Così, se il campione delle forze di lavoro deve garantire le stime sulla disoccupazione a livello regionale oltre che nazionale, per valutare l'idoneità dell'attuale numerosità campionaria occorre analizzare gli errori standard delle stime del numero dei disoccupati (o del tasso di disoccupazione) a livello regionale. Solo sulla base di tale analisi è possibile dire se il campione è sovradimensionato o sottodimensionato.

Un'ultima considerazione riguarda l'intervento del prof. Marbach. Sono d'accordo con lui quando auspica l'utilizzazione di campioni areali, ma sono dell'opinione che ciò debba essere fatto soltanto quando si sta indagando su un collettivo per il quale non si dispone di liste attendibili e complete da cui selezionare il campione.

Quello della completezza delle liste è un problema che si presenta frequentemente ed assume particolare rilevanza nelle indagini telefoniche, in cui viene coperta soltanto la popolazione con telefono mentre non sono disponibili le informazioni per la popolazione priva di telefono.

L'estensione dei risultati all'intera popolazione è possibile soltanto se sono verificate particolari ipotesi, altrimenti è necessario avvertire gli utilizzatori che i dati pubblicati sono riferiti soltanto a quella particolare popolazione.

Prima di chiudere il mio intervento devo dare risposta a due quesiti posti dalla prof.ssa Cocchi.

La prima riguarda il numero degli strati che possono essere formati utilizzando la procedura generalizzata. Il numero degli strati non è costante, ma è un parametro variabile che costituisce una opzione prevista dalla procedura.

La seconda si riferisce alla scelta del dominio territoriale (provincia, regione, etc.) entro cui si può procedere alla stratificazione dei comuni.

Per individuare l'ambito territoriale entro cui raggruppare i comuni occorre definire i domini di studio e le basi territoriali tenuto conto delle esigenze organizzative. La base territoriale entro il dominio di studio costituisce il livello territoriale utilizzato per la stratificazione.

Così se l'obiettivo è quello di fornire stime a livello regionale per l'Emilia-Romagna e dal punto di vista organizzativo si utilizzano gli uffici provinciali, allora il dominio di studio è la regione e la base territoriale la provincia. Se, invece, si ha un'organizzazione centralizzata a livello regionale, non è necessario utilizzare le province per la stratificazione. In questo caso, in genere, la stratificazione che si ottiene è più efficiente; infatti, come ha fatto osservare la prof.ssa Cocchi, si possono raggruppare comuni omogenei anche se appartenenti a province diverse.

LUIGI FABBRIS

Ringrazio il prof. Herzel e quanti altri sono intervenuti con commenti e suggerimenti sul lavoro da me svolto con la collaborazione dei Dott. Gaggiotti e Zuchegna dell'Istat e con il contributo costante dei membri della Commissione «campioni».

Rispondo alle osservazioni puntuali.

Per quanto riguarda la tecnica di selezione con probabilità variabili proposta dal Prof. Herzel, effettivamente nel rapporto si trovano giudizi discordanti in pagine diverse. Tuttavia, ciò dipende dal fatto che le elaborazioni non sono state completate e non è pertanto possibile esprimere un giudizio omogeneo. Tengo però a far presente che la mancanza non è attribuibile a chi ha scritto il rapporto, considerato che il completamento è stato più volte sollecitato, anche per iscritto. Non vado oltre per non sembrare polemico. Mi sembra tuttavia doveroso completare le elaborazioni e, in funzione dei

risultati, modificare i giudizi. Sottolineo, a scanso di equivoci, che le elaborazioni possono essere effettuate solo all'interno dell'Istat.

Per lo stimatore della varianza della varianza nel campionamento sistematico, siccome nella letteratura non esistono proposte specifiche, ne ho scelto uno asintotico per campionamenti con probabilità variabili. Chiaramente, si tratta di una soluzione di ripiego.

Per quanto riguarda l'accettabilità della definizione di stimatore basato sul quoziente da me adottata, su cui si è espresso il Prof. Cicchitelli, concordo che non si tratta dello stimatore tradizionale (che prescinde dall'ordine di estrazione delle unità), ma di uno stimatore appartenente alla classe degli stimatori basati sul quoziente che nelle elaborazioni svolte ha un ruolo preciso, in quanto ché uno stimatore che prescindesse dall'ordine di uscita delle unità, darebbe stime indifferenziate per la maggior parte delle tecniche confrontate. In ogni caso, anche se non si condizionasse il calcolo delle stime all'ordine di estrazione delle unità, la sostanza delle considerazioni svolte non cambierebbe.

Non avendo altri commenti sulle osservazioni puntuali, approfitto della tribuna per inserirmi nel dibattito culturale avviato questa mattina dalle relazioni generali.

Prima di tutto, io considero fittizia la discussione sull'essere la provincia un dominio di riferimento, un carattere di stratificazione, oppure né l'uno né l'altro. Se è un dominio di riferimento, non può — per sua natura — non essere un carattere di stratificazione. Se la provincia non è un carattere di stratificazione, il dominio di studio deve essere più ampio della provincia. Quale sia il livello territoriale del dominio, lo deve comunque definire l'analista del contenuto delle informazioni rilevate, non lo statistico.

La mia opinione è che, nello stabilire domini di studio più ampi della provincia, sia o no il dominio un carattere di stratificazione, non si può prescindere dall'ambito amministrativo nel formare il campione. Per esempio, se l'ambito territoriale supera in dimensione la provincia, ma è interno alla regione, la regione resta comunque un ambito di stratificazione. Inoltre, io non vedrei la possibilità di saltare il comune come unità di base nella formazione del campione, se non aggregando comuni contigui troppo piccoli e spezzettando comuni troppo grandi al fine di formare aree di dimensione meno variabile delle attuali.

Tra i motivi per cui, nel formare campioni sul territorio, manterrei il riferimento comunale vi è l'esistenza delle liste della popolazione, che vanno valorizzate a completamento e in concorrenza con le liste di aree. Cito a questo proposito gli Stati Uniti, dove l'indagine Current Population Survey — equivalente alla nostra indagine sulle Forze di Lavoro — si è svolta per molti anni su campioni di aree. Nel 1978, il disegno di campionamento è stato rivisto utilizzando sia liste nominative che aree. La base di indirizzi è stata formata cumulando le informazioni rilevate in varie occasioni d'indagine. A me sembra che da ciò derivi la considerazione che la scelta del disegno ottimo per una indagine non è un problema astratto, ma di concreta utilizzazione delle informazioni disponibili. Se, come spesso si verifica, le liste da cui il campione è tratto non sono accurate (supponiamo, tanto per esemplificare, che per svolgere un'indagine sulla popolazione presente si disponga delle liste di quella residente), si deve far ricorso ad altre fonti informative, e per questo si parla di campionamenti di aree concorrenti con quelli da liste.

Per quanto riguarda la numerosità campionaria, legherei il discorso al numero di strati. Come ricordava questa mattina il Dott. Zannella, nello stratificare i comuni per formare il campione per l'indagine sulle forze di lavoro, l'efficienza delle stime è quasi indifferente dalla soluzione di stratificazione adottata. Egli ha confrontato varie ipotesi di stratificazione, quella attualmente adottata, una che tiene conto solo della dimensione dei comuni e altre basate su tecniche multivariate (che tengono conto di tutte le informazioni disponibili), come la *cluster analysis*, e ha dimostrato che, alla fin fine, anche le tecniche di stratificazione più sofisticate non danno risultati più soddisfacenti di quella attuale. La soluzione apparentemente più banale, quella per dimensione, ottenuta ordinando i comuni dal più grande al più piccolo, dà risultati ottimi per tutte le stime legate alla dimensione dei comuni, ossia per quasi tutte le stime per cui si effettua attualmente l'indagine.

Per mio conto, ho effettuato una serie di sperimentazioni sugli stessi comuni sui quali abbiamo condotto le prove presentate stamani. Una frazione variabile tra il 75 e l'85%, e in alcuni casi anche il 95% della varianza è assorbito dai comuni sopra i 20 mila abitanti. La stratificazione per dimensione è dunque efficiente perché sono i comuni più grandi ad assorbire la parte più consistente di variabilità. Allora la scelta del numero strati dipende dal modo in cui si organizza la stratificazione: stratificare per dimensione implica anche sta-

bilire la soglia sopra la quale conviene inserire il comune nel campione con certezza (probabilità uno). Va detto che nelle elaborazioni svolte dalla Commissione i comuni sopra i 20.000 abitanti non sono mai stati considerati.

Un'ultima riflessione sulla numerosità del campione di comuni dentro gli strati formati. Innanzitutto, penso sia appropriato restringere la scelta tra un solo comune o due comuni per strato. Per più di due comuni, non trovo utilità se non teoriche e in casi particolari. Io sono convinto che convenga selezionare due comuni per strato; i motivi li ho già esposti nel rapporto scritto. Sono contento di sentire che la quasi totalità di quelli che si sono espressi a tal proposito concorda su questa scelta.

Concludo con una considerazione che si colloca solo in parte in questa giornata dedicata al campionamento statistico. Io credo che nelle indagini dell'Istituto Centrale di Statistica si debba passare dalla logica attuale centrata sul disegno di campionamento ad una basata sulla ottimizzazione dell'intero disegno di rilevazione, e di cui il disegno di campionamento è una parte.

Le indagini che l'Istat svolge riguardano per buona parte la rilevazione diretta di persone e di famiglie, spesso impiegando intervistatori. I rispondenti commettono errori che mettono a repentaglio l'attendibilità delle stime. Gli errori dei rilevatori vanno a cumularsi con quelli dei rispondenti, e non possono essere ignorati sia nel calcolo della variabilità delle stime, sia nel progettare il disegno di rilevazione. L'idea del controllo dell'errore di rilevazione va inserita nella scelta del disegno e nella procedura di selezione delle unità. Vanno dunque previsti, oltre al numero ottimo di comuni, anche la loro resistenza a rispondere, il numero ottimale di rilevatori per area, nonché il numero ottimale di interviste per rilevatore o la superficie media da percorrere se i rilevatori viaggiano su aree vaste.

Tutti questi elementi vanno tenuti in conto nel formulare il disegno di rilevazione. Tener conto dell'errore globale di rilevazione vuol dire — lo ripeto — impostare la rilevazione in modo da riuscire a misurarlo e, sapendo quanto grava sulle stime, manovrarlo, distribuendo il campione tra le componenti in modo appropriato. Pur avendo sull'argomento niente più che la sensazione viva della necessità, credo che quello del lavorare nella direzione della ottimizzazione del disegno di rilevazione sia un obiettivo finale per una commissione che si proponga lo studio di metodiche per la produzione di stime

di qualità. Se la cultura delle rilevazione per campione è basilare per la formazione dello statistico, quella del dominio dell'errore statistico globale è una cultura ben più completa e fruttuosa per la diffusione della statistica e per la produzione di statistiche.

GUALTIERO SCHIRINZI

Dato il ristretto tempo a mia disposizione, mi limiterò a prendere atto, con soddisfazione, dell'aver dato spazio in questa giornata di studio, oltre che al campionamento anche al problema della qualità dei dati.

Nel settore dell'agricoltura, al quale io mi dedico, i due aspetti sono fortemente sentiti. Innanzitutto quello della qualità dei dati, in quanto, trovandoci alle soglie di un censimento dell'agricoltura, esso si propone in forma preminente.

Il questionario che sarà adottato per il prossimo censimento contiene un ampio ventaglio di notizie dovuto in parte alle richieste comunitarie ed in parte all'allargamento delle tematiche sollecitate in campo nazionale. Saranno rilevati con un'analisi abbastanza ampia sia i fattori produttivi impiegati in agricoltura che fenomeni di carattere sociale legati alla struttura aziendale. A ciò occorre aggiungere che i dati individuali relativi alle singole aziende agricole, opportunamente resi anonimi, verranno forniti agli Enti territoriali nazionali (Regioni, Province, Comuni) ed all'Istituto statistico delle Comunità Europee (Progetto Eurofarm).

In queste condizioni, le indagini di qualità per il censimento devono assumere il loro necessario ruolo, per garantire ai dati prodotti un alto grado di significatività. Il programma previsto per il prossimo censimento, amplia il quadro delle indagini del censimento del 1982 limitate a due: una, eseguita col metodo della reintervista, ha riguardato l'operato dei rilevatori e lo studio del comportamento di alcune distribuzioni dei caratteri rilevati (eseguita in collaborazione col dott. Masselli) e l'altra la fase di registrazione e trattamento informatico dei dati.

L'altro aspetto è quello dei campioni in agricoltura. I campioni adottati finora sono stati sempre basati su una stratificazione delle aziende agricole su dati di carattere fisico (forma di conduzione, superficie totale, superficie agricola utilizzata, ecc.). Così è avvenuto

per le indagini sulla struttura delle aziende agricole del 1967 in poi e per le indagini delle statistiche correnti. L'adozione in campo comunitario di una classificazione tipologica delle aziende agricole mediante dati di natura economica, consente raggruppamenti delle aziende più omogenei di quanto possa realizzarsi con elementi fisici.

Uno studio del comportamento dei campioni stratificati in base all'orientamento tecnico-economico delle aziende rispetto a quelli di natura fisica, è previsto nell'ambito del vasto programma di ristrutturazione delle statistiche agricole sollecitata anche in campo comunitario. Questo potrebbe portare da una parte ad un guadagno di informazione e dall'altra una diminuzione del numero di aziende da intervistare con conseguente risparmio di risorse.

Questa giornata di studio può, in merito, costituire un autorevole stimolo per accentrare l'attenzione anche sugli aspetti che ho brevemente evidenziato necessari per consentire un ulteriore sviluppo delle statistiche agricole.

«PUBBLICAZIONI ISTAT»

BOLLETTINO MENSILE DI STATISTICA
La più completa ed autorevole raccolta di dati congiunturali concernenti l'evoluzione dei fenomeni demografici, sociali, economici e finanziari
Abbonamento annuo L. 115.000 (Estero L. 139.000) Ogni fascicolo L. 15.000
INDICATORI MENSILI
Forniscono dati riassuntivi e tempestivi sull'andamento mensile dei principali fenomeni interessanti la vita nazionale
Abbonamento annuo L. 29.000 (Estero L. 35.000) Ogni fascicolo L. 3.700
NOTIZIARI ISTAT
Forniscono i primi risultati delle rilevazioni ed elaborazioni statistiche riguardanti l'attività produttiva, i prezzi, il commercio interno, gli scambi internazionali come pure lo stato ed il movimento della popolazione e le sue caratteristiche sociali e sanitarie.
I dati, esposti in grafici e tabelle, sono accompagnati da commenti, illustrazioni e note interpretative.
Serie 1 - Statistiche demografiche e sociali
Abbonamento annuo L. 22.000 (Estero L. 29.000) una copia L. 1.600
Serie 2 - Statistiche dell'attività produttiva
Abbonamento annuo L. 64.000 (Estero L. 85.000) una copia L. 1.600
Serie 3 - Statistiche del lavoro, delle retribuzioni e dei prezzi
Abbonamento annuo L. 22.000 (Estero L. 29.000) una copia L. 1.600
Serie 4 - Argomenti vari
Abbonamento annuo L. 13.000 (Estero L. 17.000) una copia L. 1.600
Abbonamento annuo a tutte le serie L. 106.000 (Estero L. 144.000).
INDICATORI TRIMESTRALI
Conti economici trimestrali
Abbonamento annuo L. 11.000 (Estero L. 13.000) Ogni fascicolo L. 3.700
STATISTICA DEL COMMERCIO CON L'ESTERO
Documentazione statistica ufficiale, a periodicità trimestrale, sul commercio dell'Italia con l'estero; fornisce, per tutte le merci comprese nella classificazione merceologica della tariffa dei dazi doganali, l'andamento delle importazioni e delle esporazioni da e per i principali Paesi
Abbonamento annuo L. 99.000 (Estero L. 112.000) Ogni fascicolo L. 31.000

Abbonamento annuo cumulativo a tutti i periodici, compresa la "Statistica del commercio con l'estero": L. 300.000 (Estero L. 390.000); esclusa la "Statistica del commercio con l'estero" L. 209.000 (Estero L. 286.000)
Gli abbonamenti decorrono dal 1° gennaio anche se sottoscritti nel corso dell'anno. In tal caso l'abbonato riceverà i numeri dell'annata già pubblicati. L'abbonato ai periodici ISTAT ha diritto a ricevere gratuitamente i fascicoli non pervenutigli soltanto se ne segnalerà il mancato arrivo entro 10 giorni dal ricevimento del fascicolo successivo. Decorso tale termine, si spediscono solo contro rimessa dell'importo. Le variazioni di indirizzo devono essere segnalate dall'abbonato per iscritto. Nel sottoscrivere l'abbonamento cumulativo, gli interessati possono chiedere che l'ISTAT provveda, senza ulteriori richieste, all'invio di tutte le pubblicazioni non periodiche non appena liberate dalle stampe, contro assegno o con emissione di fattura, con lo sconto del 30%. Le singole pubblicazioni possono essere richieste direttamente all'Istituto nazionale di statistica (Via Cesare Balbo, 16 - 00100 Roma) versando il relativo importo, maggiorato del 10% per spese di spedizione, sul c/c postale n. 619007.
Tutti i prezzi sono riferiti all'anno 1991.

ANNUARIO STATISTICO ITALIANO - Edizione 1990 - L. 46.000
Sintetizza in semplici tabelle numeriche di facile lettura ed attraverso appropriate note illustrative e rappresentazioni grafiche, i dati fondamentali della vita economica, demografica e sociale e fornisce un quadro panoramico della corrispondente situazione degli altri principali Paesi del mondo.
COMPENDIO STATISTICO ITALIANO - Edizione 1990 - L. 22.000
Sintetizza i risultati delle rilevazioni ed elaborazioni statistiche di maggior interesse nazionale.
ITALIAN STATISTICAL ABSTRACT - Edition 1990 - L. 22.000
Fornisce i principali risultati delle rilevazioni ed elaborazioni statistiche concernenti la situazione sociale ed economica italiana - Edizione in lingua inglese.
I CONTI DEGLI ITALIANI - Vol. 24, edizione 1990 - L. 16.000
Illustra in forma divulgativa i principali aspetti quantitativi dell'economia italiana.
LE REGIONI IN CIFRE - Edizione 1991 - Distribuzione gratuita *(in corso di stampa)*
Fornisce i dati delle singole regioni e delle due grandi ripartizioni geografiche: Nord-Centro e Mezzogiorno.

ANNUARI

STATISTICHE DEMOGRAFICHE
n. 34 - Anno 1985
Tomo 1, parte prima - Movimento e calcolo della popolazione secondo gli atti anagrafici - L. 11.000
Tomo 1, parte seconda - Iscrizioni e cancellazioni anagrafiche per trasferimento di residenza, 1984 - Espatriati e rimpatriati, 1985 - L. 9.500
n. 33/34 - Anni 1984 e 1985
Tomo 2, parte prima - Nascite e decessi - L. 38.000
Tomo 2, parte seconda - Matrimoni, separazioni e divorzi - L. 15.000
n. 35 - Anno 1986
Tomo 1, parte prima - Popolazione residente e movimento anagrafico dei Comuni - L. 11.500
Tomo 1, parte seconda - Iscrizioni e cancellazioni anagrafiche, 1985 e 1986 - Espatriati e rimpatriati, 1986 - L. 15.800
n. 36 - Anno 1987
Tomo 1, parte prima - Popolazione residente e movimento anagrafico dei Comuni - L. 18.900
Tomo 1, parte seconda - Iscrizioni e cancellazioni anagrafiche - Espatri e rimpatri, 1987 - L. 15.000
n. 35/36 - Anni 1986 e 1987
Tomo 2, parte prima - Nascite e decessi *(in preparazione)*
Tomo 2, parte seconda - Matrimoni, separazioni e divorzi - L. 16.000 *(in corso di stampa)*
Raccoglie i dati sulla dinamica demografica italiana, sia naturale che migratoria, nonchè dei dati sintetici sul movimento annuale della popolazione residente anagrafica comunale e sul suo ammontare.

POPOLAZIONE E MOVIMENTO ANAGRAFICO DEI COMUNI - n. 2 - Anno 1989 - L. 20.000
Riporta i dati dell'ammontare della popolazione residente, desunti dall'analisi del movimento naturale e di quello migratorio, nonchè la stima della popolazione residente per sesso ed età a livello regionale.

STATISTICHE DELLA SANITA' - n. 3 - Anno 1987 - L. 23.000
Riunisce le statistiche sulle strutture e sull'attività degli Istituti di cura, sulle malattie infettive e diffusive soggette a denuncia obbligatoria, sulle interruzioni volontarie della gravidanza e sugli aborti spontanei.

CAUSE DI MORTE - n. 3 - Anno 1987 - L. 25.000
Raccoglie i dati relativi alle statistiche sulle cause di morte e di nati-mortalità.

STATISTICHE DELLA PREVIDENZA, DELLA SANITA' E DELL'ASSISTENZA SOCIALE
n. 29 - Anni 1988, 1989 - L. 22.000
Vengono illustrate alcune forme di attività svolte dai vari Istituti nel settore della previdenza sociale, i conti economici delle Unità Sanitarie Locali e degli Istituti ospedalieri pubblici, nonchè i principali aspetti dell'assistenza sociale.

STATISTICHE DELL'ISTRUZIONE - n. 40 - Anno scolastico 1986-87
Tomo 1 - Dati analitici: nazionali, regionali e provinciali - L. 23.000
Tomo 2 - Dati riassuntivi comunali - L. 18.000
Quadro statistico completo ed aggiornato della situazione scolastica del Paese, attraverso dati sui vari rami d'insegnamento esaminati sotto i più interessanti aspetti dell'ordinamento degli studi e dei risultati conseguiti dagli iscritti.

STATISTICHE CULTURALI - n. 29 - Anno 1987 - L. 14.000
Documentazione ufficiale completa sulle principali attività culturali concernenti, tra l'altro, la produzione libraria, la pubblicazione di riviste scientifiche, la stampa periodica e le biblioteche.

STATISTICHE GIUDIZIARIE - n. 36 - Anno 1988 - L. 41.000
Ampia documentazione statistica dell'attività giudiziaria nonchè dei principali fenomeni in materia civile e penale nel campo della criminalità e degli Istituti di prevenzione e pena.

STATISTICHE DELL'AGRICOLTURA, ZOOTECNIA E MEZZI DI PRODUZIONE - n. 36 - Anno 1988 - L. 41.000
Contiene i dati relativi ai vari aspetti dell'agricoltura nazionale, nonchè i dati sulla consistenza e produttività degli allevamenti.

STATISTICHE FORESTALI - n. 41 - Anno 1988 - L. 16.000 *(in corso di stampa)*
Fornisce un quadro completo sulla struttura delle foreste italiane e delle relative utilizzazioni legnose, unitamente ad alcuni aspetti economici.

STATISTICHE METEOROLOGICHE - n. 24 - Anno 1983 - L. 15.800
Raccoglie i dati relativi alle temperature, piovosità e altri fattori climatici rilevati da una rete di stazioni ed osservatori distribuiti nel territorio nazionale.

STATISTICHE DELLA CACCIA E DELLA PESCA - n. 4 - Anno 1988 - L. 12.000 *(in corso di stampa)*
Raccoglie i dati sull'attività della pesca e sulla consistenza del relativo naviglio, nonchè su alcuni aspetti del settore venatorio.

STATISTICHE INDUSTRIALI - n. 28 - Anni 1986 e 1987 - L. 41.000
Nel suo genere, unica e veramente preziosa pubblicazione in cui sono organicamente raccolte tutte le informazioni statistiche fondamentali concernenti il complesso ed importante settore dell'industria.

STATISTICHE DELL'ATTIVITA' EDILIZIA - n. 2 - Anno 1987 - L. 14.000
Fornisce i risultati del settore dell'attività edilizia relativamente ai fabbricati residenziali e non residenziali.

STATISTICHE DELLE OPERE PUBBLICHE - n. 2 - Anno 1987 - L. 10.000
Statistica ufficiale delle opere pubbliche effettuate dallo Stato e da Enti pubblici, nonchè da privati con finanziamento parziale dello Stato.

STATISTICHE DEL COMMERCIO INTERNO - n. 30 - Anni 1987, 1988 - L. 15.000
Fornisce i risultati delle rilevazioni correnti relativi al fenomeno della distribuzione. Vi figurano gli indici mensili delle vendite al minuto, nonchè la più recente distribuzione per Comune delle licenze di esercizio.

STATISTICHE DEL TURISMO - n. 4 - Anno 1989 - L. 12.000 *(in corso di stampa)*
Descrive il sistema delle informazioni statistiche sul turismo ed espone, in un quadro organico, statistiche, dati ed indicatori aventi per oggetto i principali aspetti di questo fenomeno.

STATISTICHE DELLA NAVIGAZIONE MARITTIMA - n. 43 - Anno 1988 - L. 20.000
Contiene i dati statistici sul movimento dei natanti e del relativo carico avvenuto nei porti marittimi e negli altri approdi autorizzati del territorio nazionale.

STATISTICA DEGLI INCIDENTI STRADALI - n. 37 - Anno 1989 - L. 20.000
La più completa ed aggiornata raccolta di dati su una materia di viva attualità.

STATISTICA ANNUALE DEL COMMERCIO CON L'ESTERO - n. 44 - Anno 1987
Tomo 1 - Dati generali e riassuntivi - L. 41.000
Tomo 2 - Merci per Capitoli merceologici e Paesi
- Parte prima: da Cap. 1 a Cap. 24 - L. 14.000
- Parte seconda: da Cap. 25 a Cap. 40 - L. 18.000
- Parte terza: da Cap. 41 a Cap. 67 - L. 21.000
- Parte quarta: da Cap. 68 a Cap. 83 - L. 18.000
- Parte quinta: da Cap. 84 a Cap. 85 - L. 25.000
- Parte sesta: da Cap. 86 a Cap. 99 - L. 18.000
- Appendice: L. 10.000

Riporta i dati definitivi sull'andamento delle importazioni e delle esportazioni con l'analisi completa del movimento per merci e per Paesi. Nel tomo primo è riportata, tra l'altro, un'ampia documentazione sul movimento delle merci nei depositi doganali e sul commercio di transito.

STATISTICHE DEI BILANCI DELLE AMMINISTRAZIONI REGIONALI, PROVINCIALI E COMUNALI - n. XXVII - Anno 1982 - L. 14.000
Espone i dati relativi ai bilanci delle Amministrazioni, tenendo conto dell'aspetto contabile, funzionale ed amministrativo dei documenti contabili. Per le Amministrazioni provinciali e comunali è stata dedicata particolare attenzione ai dati riguardanti i servizi sociali, i settori d'intervento nel campo economico ed il personale.

STATISTICHE DEL LAVORO - n. 26 - Anno 1984 - L. 12.000
Organica ed aggiornata documentazione statistica su tutti i principali aspetti del mondo del lavoro.

CONTABILITA' NAZIONALE - n. 15 - Anni 1960-85 - L. 17.000
Contiene i dati sulla struttura e sulla evoluzione delle principali grandezze del sistema economico italiano.

COLLANA D'INFORMAZIONE

Anno 1991

n. 1 - CONTI ECONOMICI DELLE IMPRESE CON 20 ADDETTI ED OLTRE - Anno 1988 - L. 22.000
n. 2 - GLI IMPIEGHI DELL'ENERGIA IN ITALIA NEL 1985 - L. 22.000
n. 3 - CONTI ECONOMICI DELLE IMPRESE PUBBLICHE CON 20 ADDETTI ED OLTRE - Anni 1983-87 - L. 12.000
n. 4 - RILEVAZIONE DELLE FORZE DI LAVORO - Gennaio 1990 - L. 12.000 *(in corso di stampa)*
n. 5 - LA SUPERFICIE FORESTALE NELLE COMUNITA' MONTANE AL 31 DICEMBRE 1989 - L. 12.000
n. 6 - CONTI DELLE AMMINISTRAZIONI PUBBLICHE E DELLA PROTEZIONE SOCIALE - Anni 1984-89 - L. 16.000
n. 7 - RILEVAZIONE DELLE FORZE DI LAVORO - Aprile 1990 - L. 12.000
n. 8 - CONTI NAZIONALI ECONOMICI E FINANZIARI DEI SETTORI ISTITUZIONALI - Anni 1980-88 - L. 16.000
n. 9 - STATISTICHE DELLA COOPERAZIONE AGRICOLA - Anno 1988 - L. 12.000
n. 10 - STATISTICHE DEL MOVIMENTO DELLA NAVIGAZIONE NEI PORTI ITALIANI - Anno 1989 - L. 12.000
n. 11 - STATISTICHE DELL'ISTRUZIONE - Dati sommari dell'anno scolastico 1988-89 - L. 22.000
n. 12 - STATISTICHE DELL'ISTRUZIONE - Dati sommari dell'anno scolastico 1989-90 - L. 22.000
n. 13 - COMMERCIO, ALBERGHI E SERVIZI VARI PER COMUNE AL 31 DICEMBRE 1988 - L. 22.000 *(in corso di stampa)*
n. 14 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (CAMPANIA) - L. 12.000 *(in corso di stampa)*
n. 15 - LAVORO E RETRIBUZIONI - Anno 1989 - L. 12.000 *(in corso di stampa)*
n. 16 - STATISTICHE DELLA ZOOTECNIA E DEI MEZZI DI PRODUZIONE IN AGRICOLTURA - Anno 1989 - L. 12.000
n. 17 - CONTI ECONOMICI DELLE IMPRESE CON ADDETTI DA 10 A 19 - Anno 1988 - L. 12.000 *(in corso di stampa)*
n. 18 - ACQUEDOTTI E RETI DI DISTRIBUZIONE DELL'ACQUA POTABILE IN ITALIA - Anno 1987 - L. 22.000 *(in corso di stampa)*

NOTE E RELAZIONI

Anno 1989

n. 1 - MANUALE DI TECNICHE DI INDAGINE (n. 7 fascicoli)
1. Pianificazione della produzione dei dati - L. 10.000 - 2. Il questionario: progettazione, redazione e verifica - L. 11.000 - 3. Tecniche di somministrazione del questionario - L. 11.000 - 4. Tecniche di campionamento: teoria e pratica - L. 20.000 - 5. Tecniche di stima della varianza campionaria - 6. Il sistema di controllo della qualità dei dati *(in corso di stampa)* - 7. Le rappresentazioni grafiche di dati statistici - L. 15.000
n. 2 - DISTRIBUZIONE PER ETA' DELLA POPOLAZIONE SCOLASTICA - Anno scolastico 1984-85 - L. 10.000
n. 3 - LA CRIMINALITA' ATTRAVERSO LE STATISTICHE - Anni 1971-87 - L. 14.000
n. 4 - PREVISIONI DELLA POPOLAZIONE RESIDENTE PER SESSO, ETA' E REGIONE - Base 1-1-1988
Tomo 1 - L. 18.000
Tomo 2 - L. 38.000
n. 5 - STATISTICHE SUI MINORENNI - Anni 1984-86 - L. 18.000
n. 6 - ANALISI DELLE FONTI STATISTICHE PER LA MISURA DELL'IMMIGRAZIONE STRANIERA IN ITALIA: ESAME E PROPOSTE - L. 10.000
n. 7 - NUMERI INDICI DEI PREZZI ALLA PRODUZIONE DEI PRODOTTI INDUSTRIALI - Base 1980 = 100 - L. 10.000

Anno 1990

n. 1 - METODOLOGIA E ANALISI DEI RISULTATI DELL'INDAGINE SULLE COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 - L. 11.000
n. 2 - LA MORTALITA' DIFFERENZIALE SECONDO ALCUNI FATTORI SOCIO-ECONOMICI - Anni 1981-82 - L. 11.000

METODI E NORME

Serie A

n. 18 - NUMERI INDICI DEL COSTO DI COSTRUZIONE DI UN FABBRICATO RESIDENZIALE: Base 1976 = 100 - L. 1.500
n. 20 - NUMERI INDICI DEI PREZZI: Base 1980 = 100 - L. 4.500
n. 21 - NUMERI INDICI DEI PREZZI DEI PRODOTTI VENDUTI E DEI BENI ACQUISTATI DAGLI AGRICOLTORI: Base 1980 = 100 - L. 5.000
n. 23 - NUMERI INDICI DEI PREZZI AL CONSUMO: Base 1985 = 100 - L. 6.300
n. 25 - NUMERI INDICI DELLA PRODUZIONE INDUSTRIALE: Base 1985 = 100 - L. 11.000
n. 26 - NUMERI INDICI DEI PREZZI ALLA PRODUZIONE DEI PRODOTTI INDUSTRIALI: Base 1980 = 100 - L. 11.000
n. 27 - NUMERI INDICI DEL FATTURATO, DEGLI ORDINATIVI E DELLA CONSISTENZA DEGLI ORDINATIVI: Base 1985 = 100 - L. 11.000

Serie B

n. 21 - ISTRUZIONI PER LA RILEVAZIONE STATISTICA DEL MOVIMENTO DELLA POPOLAZIONE - L. 4.000
n. 22 - ISTRUZIONI PER LA RILEVAZIONE DEI DATI DELLE STATISTICHE FORESTALI - L. 6.000
n. 23 - ISTRUZIONI PER LA RILEVAZIONE DELL'ATTIVITA' EDILIZIA - L. 8.400
n. 24 - ISTRUZIONI PER LE RILEVAZIONI DELLE STATISTICHE GIUDIZIARIE
Tomo 1 - Procedura di rilevazione - L. 15.800
Tomo 2 - Modelli di rilevazione - L. 15.800
n. 25 - MANUALE PER LA PROGETTAZIONE DEI DATI STATISTICI - L. 10.000
n. 26 - ISTRUZIONI PER LE COMMISSIONI COMUNALI DI CONTROLLO DELLE RILEVAZIONI DEI PREZZI AL CONSUMO - L. 10.000
n. 27 - ISTRUZIONI PER LA RILEVAZIONE DELLE OPERE PUBBLICHE - L. 11.000
n. 28 - ISTRUZIONI PER LA RILEVAZIONE STATISTICA DEGLI INCIDENTI STRADALI - L. 11.000

Serie C

n. 8 - CLASSIFICAZIONE DELLE ATTIVITA' ECONOMICHE - Edizione 1981 - L. 6.500
n. 9 - CLASSIFICAZIONE DELLE PROFESSIONI - Edizione 1981 - L. 6.500
n. 10 - CLASSIFICAZIONI DELLE MALATTIE, TRAUMATISMI E CAUSE DI MORTE - Ristampa 1986
Vol. 1: Introduzione e parte sistematica - L. 16.000
Vol. 2: Indici alfabetici - L. 25.000

ANNALI DI STATISTICA

Serie IX

Vol. 1 - ATTI DEL 2° CONVEGNO SULL'INFORMAZIONE STATISTICA IN ITALIA (Roma, 17-19 giugno 1981) - L. 10.000
Vol. 3 - STUDI STATISTICI SUI CONSUMI - Dati dal 1959 al 1974 - L. 9.500
Vol. 5 - ATTI DEL SEMINARIO SULLA VALUTAZIONE DEI RISULTATI E DELLA METODOLOGIA DEI CENSIMENTI (Roma, 7-11 maggio 1984) - L. 25.000
Vol. 6 - ATTI DEL CONVEGNO "LA FAMIGLIA IN ITALIA" (Roma, 29-30 ottobre 1985) - L. 14.000
Vol. 7 - ATTI DEL CONVEGNO SULL'INFORMAZIONE STATISTICA E I PROCESSI DECISIONALI (Roma, 11-12 dicembre 1986) - L. 15.000
Vol. 8 - ATTI DEL SEMINARIO SULLE STATISTICHE ECOLOGICHE (Roma, 28 marzo-1 aprile 1988) - L. 23.000
Vol. 9 - NUOVA CONTABILITA' NAZIONALE - L. 23.000
Vol.10 - ATTI DELLA GIORNATA DI STUDIO SUL CAMPIONAMENTO STATISTICO (Roma, 27 Aprile 1989) - L. 25.000 - *(in corso di stampa)*

CENSIMENTI

12° CENSIMENTO GENERALE DELLA POPOLAZIONE - 25 ottobre 1981
DATI SULLE CARATTERISTICHE STRUTTURALI DELLA POPOLAZIONE E DELLE ABITAZIONI - Campione al 2% dei fogli di famiglia - Dati provvisori - L. 5.000
Vol. I - Primi risultati provinciali e comunali sulla popolazione e sulle abitazioni *(dati provvisori)* - L. 6.500

Vol. II - Dati sulle caratteristiche strutturali della popolazione e delle abitazioni:
Tomo 1 - Fascicoli provinciali - Prezzi vari
Tomo 2 - Fascicoli regionali - Prezzi vari
Tomo 3 - Fascicolo nazionale - Italia - L. 25.000
Vol. III - Popolazione delle frazioni geografiche e delle località abitate dei comuni - Fascicoli regionali e nazionale - Prezzi vari
Vol. IV - Atti del censimento - L. 26.500
Vol. V - Relazione generale sul censimento - L. 25.000

POPOLAZIONE LEGALE DEI COMUNI - L. 8.000

6° CENSIMENTO GENERALE DELL'INDUSTRIA, DEL COMMERCIO, DEI SERVIZI E DELL'ARTIGIANATO - 26 ottobre 1981
Vol. I - Primi risultati sulle imprese e sulle unità locali - Dati provvisori
Tomo 1 - Dati nazionali, regionali e provinciali *(esaurito)*
Tomo 2 - Dati comunali *(esaurito)*
Vol. II - Dati sulle caratteristiche strutturali delle imprese e delle unità locali
Tomo 1 - Fascicoli provinciali - Prezzi vari
Tomo 2 - Fascicoli regionali - Prezzi vari
Tomo 3 - Fascicolo nazionale - Italia - L. 14.000
Vol. III - Atti del censimento - L. 11.000
Vol. IV - Relazione generale sul censimento - L. 26.500
3° CENSIMENTO GENERALE DELL'AGRICOLTURA - 24 ottobre 1982
CARATTERISTICHE STRUTTURALI DELLE AZIENDE AGRICOLE - L. 14.000
Vol. I - Primi risultati provinciali e comunali - Dati provvisori - L. 8.000
Vol. II - Caratteristiche strutturali delle aziende agricole:
Tomo 1: Fascicoli provinciali - Prezzi vari
Tomo 2: Fascicoli regionali - Prezzi vari
Tomo 3: Fascicolo nazionale - Italia - L. 11.000
Vol. III - Atti del censimento - L. 33.500
TIPOLOGIA DELLE AZIENDE AGRICOLE - Campione al 10% dei questionari d'azienda - L. 6.000
INDAGINE SULLE SUPERFICI A VITE
Vol. I - Caratteristiche delle aziende con vite
Tomo 1: Dati provinciali, regionali e nazionali - L. 33.500
Tomo 2: Dati comunali - L. 15.000
Vol. II - Caratteristiche dei vitigni - L. 33.500
L'ITALIA DEI CENSIMENTI - L. 10.000

ALTRE

INFORMAZIONE STATISTICA - Parliamone con l'ISTAT - Edizione 1988 - L. 12.000
CONOSCERE L'ITALIA - INTRODUCING ITALY - Edizione 1991 - Distribuzione gratuita *(in corso di stampa)*
SOMMARIO DI STATISTICHE STORICHE - 1926-1985 - L. 35.000
ATLANTE STATISTICO ITALIANO 1988 - L. 50.000
COMUNI, COMUNITA' MONTANE, REGIONI AGRARIE AL 31 DICEMBRE 1988 - Edizione 1990 - L. 20.000
ELENCO DEI COMUNI AL 31 DICEMBRE 1990 - Edizione 1991 - L. 16.000
STATISTICHE AMBIENTALI - Vol. I, 1984 - L. 9.000 *(esaurito)*
POPOLAZIONE RESIDENTE E PRESENTE DEI COMUNI - Censimenti dal 1861 al 1981 - L. 14.000
SOMMARIO STORICO DI STATISTICHE SULLA POPOLAZIONE - Anni 1951-87 - L. 41.000
IMMAGINI DELLA SOCIETA' ITALIANA - Edizione 1988 - L. 30.000
SINTESI DELLA VITA SOCIALE ITALIANA - Edizione 1990 - L. 15.000
MORTALITA' PER CAUSA E UNITA' SANITARIA LOCALE - Anni 1980-82 - L. 35.000
ELEZIONI DELLA CAMERA DEI DEPUTATI E DEL SENATO DELLA REPUBBLICA, 14 giugno 1987 - L. 10.000
45 ANNI DI ELEZIONI IN ITALIA 1946-90 - Edizione 1990 - L. 20.000
IL VALORE DELLA LIRA DAL 1861 al 1982 - L. 5.000
STATISTICHE SULLA AMMINISTRAZIONE PUBBLICA - Anni 1985-87 - L. 21.000

Stampato
a S. Atto (Teramo)
dalla EDIGRAFITAL S.p.A.
Giugno 1991