UNIVERSITA' DEGLI STUDI DI PISA

Dipartimento di Statistica e Matematica Applicata all'Economia

Report n. 51

Elementi di uno schema di campionamento areale per alcune rilevazioni ufficiali in Italia

Gilberto GHILARDI

Pisa, gennaio 1992

La ricerca è stata finanziata in parte dal Ministero dell'Università e della Ricerca Scientifica e Tecnologica (fondi 60%)

INDICE

Premessa

Il lavoro ha preso lo spunto da una ricerca svolta presso il Dipartimento Statistico dell'Università di Firenze. Lo scopo principale dell'attività è stato quello di prendere in considerazione la possibilità di utilizzare una procedura di campionamento areale per la raccolta efficiente di dati statistici su alcuni fenomeni per i quali la documentazione statistica appare carente. In questo campo diversi paesi hanno maturato una notevole esperienza e l'aspetto metodologico è stato approfondito da diversi studiosi. Tuttavia va notato che il campionamento probabilistico è un argomento tipico, in cui gli aspetti applicativi e teorici sono talmente connessi, che ogni schema deve essere finalizzato alla particolare situazione empirica per la quale esso viene predisposto. Questa esigenza è particolarmente avvertita nel caso della predisposizione di un campione areale ed ha portato a maturare la convinzione dell'utilità di delineare una proposta basata sulla situazione che caratterizza il nostro paese, per quanto riguarda le esigenze di documentazione statistica e la disponibilità delle informazioni di base per le operazioni di campionamento.

1. Introduzione

L'oggetto di queste pagine è un esame della possibilità di applicazione del campione areale in Italia per la raccolta di dati statistici. Le considerazioni sono state formulate facendo soprattutto un raffronto con il procedimento attualmente seguito per la rilevazione delle forze di lavoro, anche se in questo settore di indagine i risultati conseguiti appaiono adeguati in rapporto a diverse esigenze di documentazione statistica. Tuttavia, in generale, tale confronto è utile per capire quali sono o potrebbero essere i vantaggi e gli svantaggi di una procedura di campionamento areale, prefigurando i problemi che si incontrerebbero, nel caso in cui questo schema di campionamento fosse adottato per la rilevazione di dati su diversi fenomeni, per i quali l'informazione statistica è ritenuta carente.

In passato, il problema dell'adozione di un campione di tipo areale è già stato preso in esame anche per l'Italia, ma le condizioni allora esistenti avevano indotto a ritenere che i risultati non sarebbero stati soddisfacenti, soprattutto perché le aree alle quali si doveva fare riferimento non erano sufficientemente piccole per poter essere scelte proficuamente come base per la selezione delle unità statistiche e le stime campionarie sarebbero state affette da una variabilità eccessiva. Attualmente la situazione si è modificata e si prevede che per il futuro, a partire dai dati censuari dell'anno 1991, si potrà far riferimento ad una suddivisione del territorio in tante piccole aree, che potrebbero essere prese come base per effettuare le rilevazioni campionarie rivolte alla raccolta di informazioni su diversi fenomeni collettivi.

Vale la pena precisare che in queste pagine sono riportate principalmente alcune considerazioni inerenti uno schema di campionamento per l'estrazione delle unità campionarie, prendendo in esame solo alcuni degli aspetti più importanti del disegno complessivo di indagine, quali l'ampiezza del campione, il costo della rilevazione e la qualità dei risultati. Inoltre, la trattazione dell'argomento fa spesso riferimento solo ad una variabile oggetto di studio, anche se per tutti i campi di indagine le variabili sono diverse, ma gran parte delle conclusioni possono essere ritenute valide anche per più variabili.

Circa le esperienze maturate attraverso le rilevazioni campionarie, si può dire che esse hanno fatto uso di diversi metodi e procedimenti anche nell'ambito del medesimo campo di indagine. In particolare, in Italia si è fatto prevalentemente ricorso a dei campioni, che vengono formati estraendo le unità di osservazione o di rilevazione dalle liste o dai registri disponibili, mentre in altri paesi la selezione dei campioni probabilistici è stata spesso realizzata procedendo alla stesura di elenchi delle unità statistiche, elenchi che vengono predisposti per le aree territoriali selezionate secondo una metodologia di campionamento detta areale. I motivi principali per i quali in Italia il metodo e le procedure di campionamento areale non hanno trovato frequente applicazione sono probabilmente di diversa natura; tra tali motivi ha certamente avuto un peso il fatto che i registri o gli elenchi disponibili per le unità di rilevazione erano considerati validi per il raggiungimento degli scopi delle indagini statistiche condotte e che, contemporaneamente, non esisteva la

possibilità di far riferimento a delle aree territoriali idonee ad essere utilizzate ai fini della formazione di un campione areale.

Con questo lavoro desideriamo prendere nuovamente in esame il problema della formazione di un campione areale, facendo riferimento alle condizioni che attualmente si presentano ad un operatore nel campo della statistica.

Come si è già accennato, per dare una impostazione pratica alla esposizione, si è ritenuto utile considerare uno schema di campionamento areale basato sulle informazioni territoriali attualmente disponibili, per poi fare un confronto con un altro schema di campionamento. In particolare, come termine di raffronto si è preso lo schema di campionamento usato per la rilevazione trimestrale delle forze di lavoro, per cercare i mettere in luce le differenze che ci si possono attendere in termini teorici e pratici. Tuttavia notiamo che, verosimilmente, il campo di indagine delle forze di lavoro non rappresenta il caso in cui i vantaggi derivanti dall'uso di un campione areale sono più evidenti per diverse ragioni, sulle quali avremo occasione di soffermarci nelle pagine seguenti.

In merito alla trattazione dell'argomento, dal punto di vista metodologico faremo riferimento allo schema di un campione a due stadi con stratificazione delle unità di primo stadio, ovvero allo schema che viene usato anche nella rilevazione delle forze di lavoro; successivamente, descriveremo le caratteristiche di un campione areale finalizzato ad uno studio nello stesso settore di indagine. Infine, si cercherà di mettere in evidenza le differenze che ci si possono attendere tra i risultati derivanti dall'uso dei due schemi di campionamento, in funzione delle esigenze e delle circostanze che caratterizzano il contesto in cui l'attività di documentazione statistica viene svolta.

Naturalmente, oltre questi aspetti della predisposizione di un piano ve ne sono altri, quali l'organizzazione dell'indagine, la raccolta dei dati sul campo, la loro elaborazione e gli errori non campionari; nonostante la loro importanza, ci occuperemo solo marginalmente di questi argomenti, poichè spesso essi presentano dei problemi abbastanza simili nei diversi schemi di campionamento.

2. Alcuni tratti caratteristici delle rilevazioni ufficiali campionarie

Le indagini campionarie sulle forze di lavoro sono state avviate in seguito all'esperienza maturata negli Stati Uniti e hanno avuto certamente un ruolo determinante per lo sviluppo ed il consolidamento di alcuni settori metodologici della statistica (O'Muircheartaigh e Wong, 1981). Fino dagli anni trenta, periodo al quale può farsi risalire in Europa l'uso delle indagini campionarie ufficiali sulle forze di lavoro e sulle famiglie, gli schemi di campionamento che vengono utilizzati hanno una caratteristica in comune, in quanto fanno riferimento a due o più stadi di selezione delle unità di rilevazione. Infatti, l'estrazione delle unità di rilevazione viene effettuata considerando due o più livelli di campionamento, che sono individuati raggruppando le unità stesse in base alla loro collocazione territoriale.

In altre parole, nei casi più semplici, si fa ricorso alle cosiddette unità di campionamento primarie che sono rappresentate da aree territoriali non troppo ampie, tra le quali si procede all'estrazione casuale di un certo numero di unità (talvolta dopo la loro ripartizione in strati). Tra le unità primarie selezionate si procede poi alla individuazione delle unità di rilevazione (se si hanno solo due stadi) attraverso una elencazione esaustiva (microcensimenti) oppure attraverso i registri disponibili per queste unità. Così facendo, il campione risulta formato da unità appartenenti ad un numero limitato di aree territoriali e permette di ridurre i costi per la raccolta dei dati con intervista diretta. Tuttavia, occorre notare che, ovviamente, qualora per le unità primarie estratte si proceda alla elencazione esaustiva delle unità di rilevazione, la formazione del campione non presuppone la disponibilità preventiva di una lista completa delle unità da rilevare.

Sulla base dei risultati ottenuti, si effettuano generalmente le stime dei parametri interessanti per la popolazione statistica nel suo insieme, oppure per grandi regioni territoriali.

In questo contesto, si ricorre talvolta ad una stratificazione delle unità primarie, prima della loro selezione, con l'intento di ridurre la variabilità dei risultati campionari, che spesso in gran parte è dovuta alla variabilità che caratterizza tali unità. I criteri di

stratificazione che vengono usati sono generalmente le caratteristiche socio-economiche oppure la dimensione in termini di popolazione statistica. In ogni caso, spesso si procede in modo che ogni unità di osservazione abbia la stessa probabilità di essere inserita nel campione finale e, in questo senso, il campione sia autoponderante, ovvero che l'estensione dei risultati campionari possa essere effettuata applicando a ciascuna unità campione un fattore di estensione costante.

Circa l'inviduazione delle unità statistiche, quando sono disponibili dei registri anagrafici, essa viene effettuata operando una selezione casuale da tali registri; altrimenti si procede ad un'elencazione preventiva delle unità appartenenti alle unità primarie selezionate e talvolta alla loro ripartizione in tante piccole sottoaree, dette segmenti, di uguale dimensione in termini delle unità stesse. Successivamente, viene effettuata una estrazione casuale di un determinato numero di tali unità o segmenti.

Naturalmente, i due modi di procedere hanno diverse implicazioni a seconda del contesto in cui vengono adottati. Infatti, ad esempio, il primo procedimento è generalmente meno costoso, ma la bontà dei risultati è condizionata dall'aggiornamento dei registri anagrafici delle unità di rilevazione (Redfern, 1989). Invece, il secondo procedimento comporta un'attività preliminare di elencazione delle unità appartenenti alle unità primarie selezionate, alla quale spesso fa seguito la suddivisione in gruppi o segmenti; tale attività assorbe una parte delle risorse destinate allo svolgimento dell'indagine ed è più o meno costosa a seconda della dimensione delle unità primarie. In questo secondo caso ci sembra sia appropriato il riferimento alla denominazione di campione areale e la sua applicazione prescinde dalla disponibilità preventiva di un elenco aggiornato delle unità di rilevazione.

In definitiva, si può pensare che il campione areale (area sampling) sia ottenuto facendo riferimento ad una suddivisione del territorio in cui si svolge l'indagine statistica in un certo numero di piccole aree e alla scelta casuale di una parte di queste ai fini della individuazione esaustiva delle unità di rilevazione, tra le quali selezionare quelle da inserire nel campione. Pertanto la disponibilità di piccole aree territoriali alle quali appartengono le unità di rilevazione costituisce un requisito praticamente indispensabile

per lo svolgimento delle operazioni di campionamento. Lo schema di selezione delle unità che si prefigura è a due (o più) stadi, includendo in questa dizione come caso particolare il campione a grappoli, che si ha quando tutte le unità di rilevazione appartenenti alle unità primarie selezionate casualmente vengono inserite nel campione. Secondo questa impostazione, il campione areale rientra in uno schema di campionamento classico, ma è caratterizzato da accorgimenti particolari (Kish, 1965), che sono rilevanti per la soluzione dei problemi derivanti dalla variabilità degli stimatori, ai quali si ricorre nell'ambito dell'indagine statistica.

In particolare, vale la pena di notare che la variabilità dei risultati campionari derivante dal procedimento di selezione è generalmente più elevata di quella che si avrebbe usando un campione casuale semplice (Kish e Frankel, 1974) e la differenza dipende dall'omogeneità interna alle areole o gruppi (clusters) di unità, e dalla loro dimensione. Per avere un'idea dell'influenza di questi fattori sulla variabilità dei risultati campionari e dei problemi relativi, basta pensare che se internamente i gruppi fossero omogenei, poche unità per ciascun gruppo sarebbero sufficienti per riuscire a ricavare delle stime abbastanza precise, mentre si dovrebbe tendere a selezionare un numero consistente di unità primarie o gruppi, in modo da poter ridurre l'effetto prodotto sulla variabilità delle stime dalla disomogeneità esistente tra gruppi differenti.

Per quanto riguarda la situazione italiana, ricordiamo che durante gli anni 50, periodo in cui altri paesi sperimentavano e utilizzavano degli schemi di campionamento areale, ha avuto inizio l'indagine sulle forze di lavoro secondo uno schema che prevede due stadi di selezione del campione (ISTAT, 1978). Il primo stadio è rappresentato dai comuni, mentre il secondo stadio è costituito dalle famiglie. Come è noto, una parte dei comuni vengono inclusi con certezza nel campione e, secondo una terminologia corrente, sono detti autorappresentativi, mentre la parte restante dei comuni viene ripartita in strati secondo alcuni caratteri geografici e socio-economici; successivamente, ogni strato viene sottoposto ad una estrazione casuale, selezionando un comune per ogni strato, in modo che ciascun comune abbia una probabilità di essere estratto proporzionale alla sua dimensione in termini di popolazione. Le unità da rilevare vengono selezionate all'interno

di ogni comune campione attraverso i registri anagrafici con un procedimento di estrazione sistematica, consistente nella scelta casuale di un certo numero di famiglie che è effettuata sulla base di un intervallo di ampiezza K di campionamento. La frazione comunale 1/K di campionamento viene stabilita in base alla regione di appartenenza del comune, affinchè il campione risulti abbastanza ampio da fornire stime con un errore teorico non superiore al 5%.

Da quanto è stato detto, appare evidente che lo schema di campionamento usato per l'indagine sulle forze di lavoro non può essere considerato un campione areale in senso proprio, dato che questo si basa sulla possibilità di avvalersi di un certo numero di areole con un numero piccolo e costante di unità statistiche. Tuttavia, notiamo che talvolta il termine di campione areale viene usato in senso lato anche quando le areole prescelte non hanno tutte le caratteristiche desiderate, o quando si effettua una selezione casuale dagli elenchi disponibili delle unità statistiche appartenenti alle areole da inserire nel campione.

In questo senso, il campione areale è stato oggetto di considerazioni e valutazioni per la sua possibilità di impiego anche in Italia. Infatti, ad esempio, alcune indicazioni sul tema in questione si trovano in un lavoro presentato alla XVII riunione dell'Istituto Internazionale di Statistica (Gini, 1927) e successivamente in un documento di lavoro non pubblicato (Giusti, 1973). In particolare, il Gini si era posto il problema di estrarre casualmente m circoscrizioni tra M circoscrizioni, al fine di formare un campione di n unità statistiche, tali che (a priori) una media campionaria $\bar{x}$ calcolata sulle unità campione non differisse dalla media μ della popolazione di una quantità superiore al valore prefissato ϵ con probabilità assegnata. In tale circostanza le aree prescelte furono i circondari (aree di un'ampiezza intermedia tra quelle delle province e dei comuni) e fu osservato che la variabilità delle stime risultava elevata a causa della eccessiva ampiezza delle circoscrizioni considerate.

Per quanto concerne il documento di lavoro non pubblicato, predisposto per l'Istat, la valutazione sull'opportunità dell'adozione di un campione areale nell'indagine sulle forze di lavoro metteva in evidenza il vantaggio di prescindere dalla disponibilità preventiva di liste aggiornate delle unità di osservazione, ma faceva rilevare che per l'Italia

la scelta delle sezioni di censimento come base di campionamento per la selezione delle famiglie, avrebbe portato approssimativamente alla scelta di 400 sezioni su 70.000, ovvero ad un numero eccessivamente piccolo di sezioni campione.

Recentemente, l'argomento è stato oggeto di ulteriori considerazioni (Fabbris, 1990; Marbach, 1990) ed è stata auspicata anche per l'Italia una sperimentazione del campionamento areale, ma occorre dire che per la sua adozione sussistono tuttora alcune perplessità, soprattutto quando per le unità statistiche si può fare riferimento, come base di campionamento, a dei registri che sono ritenuti adeguati in rapporto alle esigenze di documentazione statistica.

Attualmente, non vi è dubbio che nella prospettiva di ricorrere all'uso di un campione areale, diversi elementi interessanti sono mutati, in quanto le rilevazioni campionarie sono più numerose, talvolta esse non possono avvalersi di elenchi o registri sufficientemente aggiornati, mentre si prevede che sarà disponibile una suddivisione territoriale dettagliata e più adatta di quelle esistenti ad essere presa come punto di riferimento per la definizione di uno schema di campionamento.

Pertanto, è senz'altro utile esaminare con attenzione il problema, tenuto conto anche del fatto che altri paesi, quali gli Stati Uniti, il Canada, la Germania, la Spagna e la Gran Bretagna (U.S.Census Bureau, 1978; Fellegi et al., 1967; Statistics Canada, 1976; F.S.O., 1969; Espana, 1975; OPCS, 1973), hanno maturato una notevole esperienza nell'uso dei campioni areali e li utilizzano correntemente per le loro indagini periodiche.

3. Le suddivisioni del territorio in piccole aree

Dato che il campione areale presuppone l'esistenza di una suddivisione territoriale in aree sufficientemente piccole, in teoria si prospettano due possibilità: procedere alla ripartizione del territorio attraverso una rete con una maglia regolare e sufficientemente piccola, oppure fare riferimento ad una delle suddivisioni esistenti.

La prima soluzione presenta difficoltà operative abbastanza consistenti, data la limitata disponibilità per l'Italia di carte geografiche sufficientemente dettagliate, tra le

quali anche quelle in scala 1:2.000 sono spesso approssimative in relazione alla individuazione di strade, vie e numeri civici, senza considerare il fatto che, quando esistono, frequentemente non risultano aggiornate. Inoltre, per le aree territoriali definite sovrapponendo una rete a maglie regolari su un supporto cartografico adeguato, va aggiunto che risulta difficile disporre dei dati statistici utili nella predisposizione di un campione areale, che le superfici individuate dalle maglie non sono uguali, che le unità sui confini sono tanto più numerose, quanto più sono piccole tali maglie e che i confini tra aree sono spesso invisibili sul territorio.

Tra le suddivisoni esistenti del territorio e che sono potenzialmente utili per la formazione di un campione areale, la ripartizione in sezioni di censimento appare uno strumento da valutare attentamente, perché, secondo una valutazione sommaria, la dimensione delle sezioni non appare abbastanza piccola ed uniforme. In tal senso, il vantaggio di fare uso delle sezioni di censimento, deriva soprattutto dal fatto che per queste aree sono disponibili delle informazioni aggiornate in occasione dei censimenti e che il loro numero è divenuto sempre più consistente, cosicché la dimensione media è notevolmente diminuita. Data la situazione, riteniamo sia interessante cercare di stabilire se, ed in quale misura, le sezioni di censimento rappresentano uno strumento adatto per la formazione di un campione areale. Tuttavia, prima di procedere nell'esame degli aspetti metodologici e pratici del problema, vale la pena di fornire qualche indicazione sulle caratteristiche delle sezioni di censimento nelle quali verosimilmente il territorio italiano risulterà suddiviso in seguito alle operazioni censuarie del 1991. Facendo riferimento alla situazione attuale ed ai criteri che l'Istat ha fornito ai comuni per la predisposizione della documentazione necessaria nello svolgimento delle operazioni censuarie, a partire dal censimento dell'anno 1991 si può prevedere un aumento consistente delle sezioni di censimento (circa 400.000), che segue quello registrato passando dal 1971 (circa 80.000 sezioni) al 1981 (circa 150.000 sezioni). Inoltre, per quanto riguarda la loro individuazione, essa dovrebbe rispondere a diversi criteri, quali l'omogeneità interna, il contenimento del numero di unità statistiche incluse (non superiore a 300 famiglie) ed il

mantenimento della possibilità di aggregazione dei dati per consentire una comparabilità dei dati stessi nel tempo.

Più in generale, la nuova suddivisione del territorio comunale dovrebbe permettere di definire un riferimento spaziale sufficientemente particolareggiato, per procedere ad analisi statistiche articolate, che finora non è stato possibile effettuare a causa dei riferimenti territoriali ad aree troppo ampie ed internamente disomogenee.

Pertanto, l'aumento considerevole del numero di sezioni, deriva da molteplici esigenze di documentazione statistica e lascia intravedere la possibilità di ricorrere alle sezioni di censimento anche per il campionamento di tipo areale nell'ambito di alcune indagini statistiche condotte ai fini di documentazione e di studio di diversi fenomeni collettivi.

Inoltre, occorre considerare che, secondo le intenzioni, le sezioni saranno classificate secondo il tipo di zona (agricola, edificata, periferica, del centro storico e così via), fornendo alcuni presupposti interessanti per una eventuale stratificazione delle sezioni stesse in gruppi omogenei rispetto a determinati criteri.

Circa l'impiego delle altre suddivisioni territoriali esistenti (aree elettorali, aree SIP, ENEL, PPTT), esso è ostacolato soprattutto dal fatto che la superficie territoriale delle areole non è sufficientemente piccola e che la disponibilità di dati statistici aggiornati è quanto meno problematica, senza considerare che talvolta si presentano delle difficoltà di raccordo con le delimitazioni esistenti di natura amministrativa.

Una ulteriore possibilità di aggregazione di dati territoriali per il momento è solo ipotetica ed è rappresentata dalla geocodifica degli indirizzi delle unità statistiche di rilevazione. Questa soluzione in teoria appare assai interessante, perché consentirebbe di esaminare e confrontare diverse aggregazioni, per poi scegliere quella più conveniente.

In pratica, riteniamo che un giudizio definitivo sui vantaggi derivanti dal ricorso alla geocodifica possa essere dato soltanto in seguito ad un studio accurato e ad una sperimentazione diretta.

Nel complesso, anche se dalle osservazioni riportate non è facile trarre elementi di giudizio definitivi sui vantaggi derivanti dall'uso di una suddivisione territoriale rispetto

alle altre, ci è sembrato interessante elaborare uno schema di campione areale basato su 400.000 sezioni di censimento, nonostante la mancanza di informazioni precise sulle loro caratteristiche.

4. Criteri per la definizione di uno schema di campionamento

Gli elementi da valutare per una definizione appropriata di uno schema di campionamento sono diversi (Kish, 1979; Redfern, 1974) e riguardano l'attività di documentazione statistica in generale, l'attività inerente il particolare settore di indagine ed i problemi da risolvere nell'ambito dell'indagine specifica. Ciò è abbastanza ovvio, se pensiamo che il sistema informativo statistico assolve ad una molteplicità di funzioni e risulta tanto più efficiente, quanto più lo svolgimento di tali funzioni è coordinato ed integrato per il raggiungimento delle finalità da perseguire.

D'altro canto, è altrettanto ovvio che è praticamente impossibile fare un bilancio completo delle connessioni esistenti tra le varie attività di documentazione, per poi programmare in maniera adeguata una determinata indagine.

Tuttavia, si può cercare di dare qualche indicazione sui criteri ai quali ci si può attenere, volendo predisporre una rilevazione statistica e tralasciando le questioni che esulano dall'ambito delle competenze dello statistico e sconfinano in altre sfere di competenza.

Una prima considerazione da fare riguarda l'opportunità di stabilire un collegamento tra rilevazioni campionarie e rilevazioni complete (Khamis e Alonzo, 1975; Kish e Verma, 1983). In particolare, si nota che dal confronto tra dati censuari e dati campionari si possono trarre diversi elementi di giudizio sulla qualità dei dati raccolti. Naturalmente, tale confronto è facilitato qualora le informazioni abbiano delle date di riferimento abbastanza prossime ed una distribuzione territoriale adeguata. Ciò, ad esempio, sarebbe abbastanza agevole qualora lo schema di campionamento fosse predisposto sulla base di dati censuari, con un criterio di tipo areale, perché in questo caso la rilevazione campionaria

sarebbe impostata in funzione di una base aggiornata (in occasione del censimento) e permetterebbe di effettuare una verifica dei dati censuari.

Infatti, la rilevazione censuaria è in grado di fornire una base di campionamento per un campione areale attraverso la definizione precisa e puntuale delle sezioni di censimento con i relativi dati statistici da impiegare per la selezione casuale delle sezioni stesse e per migliorare le stime campionarie. Inoltre, il campione areale potrebbe essere utile per controllare i dati censuari, in quanto i dati campionari possono essere raccolti mediante intervistatori preparati e con questionari più ampi rispetto a quelli usati per il censimento. In particolare, per le aree inserite nel campione sarebbe interessante effettuare un controllo accurato della copertura del censimento e, in generale, uno studio delle cause che influiscono sulla riuscita delle operazioni censuarie. Infine, i dati campionari potrebbero essere usati altrimenti per effettuare delle stime tempestive dei risultati censuari o per fornire una risposta alle esigenze che si manifestano nell'ambito dei problemi di stima per piccole aree (Drew et al., 1982; Purcell e Kish, 1980; Sarndal, 1984). Tuttavia, ad una prima riflessione, appare evidente che la base di campionamento censuaria andrebbe soggetta ad un certo invecchiamento nell'intervallo intercensuario, cosicché si renderebbe opportuna l'adozione di alcuni accorgimenti, al fine di ridurre questo inconveniente.

Una situazione abbastanza diversa si presenta quando si fa riferimento, come base di campionamento, ai registri anagrafici delle unità statistiche (Redfern, 1974 e 1979), perché questi risentono quasi sempre dei problemi di aggiornamento e rappresentano una fonte di dati di tipo amministrativo, con tutte le conseguenze che potrebbero essere riscontrate.

Il problema della scelta della base di campionamento adeguata, così come è stata posta, non sembra di facile soluzione. Mentre, se si tiene conto di alcune situazioni concrete in cui ci si trova ad operare, forse si ha un' idea più precisa dei motivi per i quali si preferisce fare una determinata scelta.

Infatti, ad esempio, nella maggioranza dei casi le rilevazioni campionarie sulle famiglie si sono basate su schemi di campionamento areale a più di due stadi di selezione, soprattutto perché non era disponibile una lista aggiornata delle unità di rilevazione.

Inoltre, in tale campo si è verificata una certa evoluzione, in quanto la base di campionamento si è arricchita di diverse informazioni e si è registrata una certa tendenza verso la diminuzione del numero degli stadi di campionamento fino al penultimo stadio, che spesso è rappresentato da unità areali sufficientemente piccole ed omogenee rispetto alla loro dimensione in termini del numero di unità statistiche.

Per quanto riguarda l'impiego di campioni selezionati a partire dalle liste anagrafiche disponibili delle unità statistiche, verosimilmente essi forniscono buoni risultati nella misura in cui tale liste sono aggiornate tempestivamente, ma si è abbastanza vincolati nell'ottenimento dei risultati alla acquisizione di informazioni sui fenomeni che riguardano le unità presenti nella lista.

Un altro aspetto interessante per valutare l'opportunità di usare un determinato schema di campionamento è quello del costo di raccolta delle informazioni per una precisione prefissata delle stime. Da questo punto di vista, almeno da una prima valutazione, l'uso di un campione areale comporta un certo sacrificio rispetto ad un campione casuale semplice, perché, com'è noto, effettuando una selezione delle unità statistiche a gruppi, si riscontra generalmente un aumento della variabilità dei risultati campionari (Kish e Frankel, 1974), un aumento che è tanto più consistente, quanto più i gruppi risultano omogenei al loro interno per le caratteristiche delle unità considerate. Perciò, a parità di precisione, il campione areale comporterebbe la rilevazione di un maggior numero di unità rispetto al campione casuale semplice e, almeno a prima vista, un maggior costo. Tuttavia, non si possono trarre in maniera così immediata delle conclusioni al riguardo, ma occorre condurre un'analisi abbastanza approfondita, effettuando un confronto con le soluzioni alternative concrete e non soltanto con un campione casuale semplice, cioè con una alternativa che spesso è solo teorica. Infatti, anche per i campioni ottenuti a partire dai registri delle unità di rilevazione quasi sempre si utilizzano più stadi di selezione, che contribuiscono ad aumentare la variabilità delle stime rispetto ad un campione casuale semplice. Inoltre, non va trascurato il fatto che si può cercare di definire uno schema di campionamento areale, in modo da farvi riferimento per più di una rilevazione statistica, conseguendo alcune economie di natura tecnica e

finanziaria e quindi diminuendo il divario esistente per la rilevazione dei dati tra i costi relativi a tale schema e quelli relativi ad altri schemi meno onerosi di campionamento.

Spesso un'analisi di questo aspetto della rilevazione statistica è limitata essenzialmente alle considerazioni sul problema della raccolta dei dati sul campo e viene condotta (Kish, 1965; Hansen et al.,1953) in relazione ad un costo C di rilevazione complessivo

$$C = C_a \frac{Ff}{g} + C_u Ff \qquad (1)$$

che è determinato in base ad un costo C_u per la raccolta dei dati presso ciascuna delle Ff unità campione tra F unità di rilevazione (con f pari alla frazione di campionamento) e ad un costo C_a da sostenere per svolgere le operazioni necessarie al fine di prendere contatto con l'unità stessa. In particolare, il costo C_a incide in maniera diversa sul costo C complessivo, in funzione del numero Ff/g dei punti di campionamento che occorre raggiungere ed individuare, qualora le Ff unità campione siano collocate sul territorio a gruppi di ampiezza g. Tale costo include le spese di viaggio per prendere contatto con l'unità da rilevare, mentre il costo C_u è da mettere in relazione soprattuto con le operazioni da svolgere per riempire materialmente i questionari individuali. Oltre il costo C per le operazioni da svolgere sul campo dell'indagine, un altro elemento interessante di uno schema a di campionamento da prendere in esame è la varianza

$$var_a(\bar{x}) = \sigma_c^2 [\, 1 + \rho\,(g - 1)] \qquad (2)$$

dello stimatore $\bar{x}$, la quale può essere scritta come una funzione della varianza σ_c^2 delle stime relative ad un campione casuale semplice e del coefficiente ρ di correlazione intraclasse. Questo coefficiente rappresenta una misura dell'omogeneità interna ai gruppi di unità da selezionare e comporta generalmente un incremento della variabilità delle stime, rispetto a quella che si avrebbe per un campione casuale semplice di unità.

Dall'esame delle due espressioni si nota che il costo complessivo C diminuisce al crescere della dimensione g dei grappoli, mentre la varianza delle stime subisce un incremento, dato che generalmente il valore del coefficiente ρ è maggiore di zero. Pertanto, è opportuno programmare la raccolta dei dati attraverso dei grappoli di una dimensione g_0, che consenta di rendere minimo il costo per una precisione prefissata oppure di ottenere la precisione massima, fissato il costo C di raccolta dei dati. Come è stato dimostrato (Hansen et al., 1953; Kish, 1965), l'espressione che consente di ricavare g_0

$$g_0 = \left[\frac{C_a}{C_u}\frac{1-\rho}{\rho}\right]^{1/2} \qquad (3)$$

dipende dal rapporto tra i costi C_a, C_u e dal coefficiente ρ di correlazione intraclasse. In proposito, l'espressione che permette di calcolare g_0, viene ricavata a partire congiuntamente dalle due espressioni riportate per il costo C e per la varianza dello stimatore $\overline{x}$.

Oltre l'opportunità del collegamento tra indagini complete ed indagini campionarie, l'esame dei costi di rilevazione e della variabilità dei risultati per un determinato schema campionario, anche il coordinamento dell'attività di documentazione relativa a più indagini campionarie riveste una certa importanza. Gli elementi che si possono individuare in questo ambito sono molteplici e si riferiscono, ad esempio, alla maggiore facilità di gestire più indagini, quando queste utilizzano lo stesso schema di selezione, oppure alla possibilità di ripartire su più rilevazioni alcuni costi comuni di raccolta dei dati. In particolare, ad esempio, l'uso parziale o totale di una medesima base di campionamento per più indagini può essere vantaggioso per diverse ragioni, tra le quali ricordiamo il fatto di far riferimento ad una struttura unica di rilevazione, riducendo alcuni costi complessivi di indagine.

Tra gli altri criteri interessanti per la messa a punto degli schemi di campionamento, viene generalmente considerato importante predisporre il campione affinchè si possa associare ad ogni stima una misura degli errori non campionari (Fellegi, 1964; Hansen et

al.,1961; Madow et al.,1983), ovvero che non sono attribuibili alla casualità del procedimento di selezione delle unità, ma ai fattori che possono aver determinato una distorsione nelle stime. Infatti, ormai è opinione abbastanza diffusa che sia necessario controllare gli effetti di tale fonte di errori, perché si ritiene che talvolta essa sia anche più importante di quella dovuta alla casualità dell'estrazione, che è implicita nell'applicazione di uno schema di campionamento.

Pertanto, dato che ogni rilevazione è affetta da errori di natura campionaria e da errori di tipo non campionario, appare evidente l'opportunità di predisporre un piano di campionamento, in modo che i dati raccolti permettano di misurare ambedue i tipi di errore (O'Muircheartaigh, 1982; Verma et al.,1980).

Infine, notiamo che da un punto di vista del tutto generale, la semplicità dello schema di campionamento viene considerata un pregio, perché ne facilita l'applicazione permettendo di operare agevolmente nella fase di raccolta dei dati ed in quella della definizione e della messa a punto dei procedimenti di stima.

Volendo fare una osservazione particolare nel caso delle rilevazioni campionarie periodiche, come è facile immaginare, esse devono avvalersi di uno schema che mantenga una certa validità nel tempo e che abbia la capacità di cogliere tempestivamente i mutamenti relativi alla situazione che caratterizza le unità di rilevazione in tempi diversi.

In questa ottica anche la flessibilità dello schema nel consentire agevolmente eventuali adeguamenti del campione, diventa un requisito importante, dato che permetterebbe di modificare il campione, ad esempio al fine di ottenere una maggiore precisione delle stime.

Al termine di questa breve riflessione sui criteri utili nel definire il procedimento di estrazione di un campione, desideriamo sottolineare che la programmazione fatta a tavolino dei compiti da svolgere può senz'altro tracciare le direttive secondo le quali occorre lavorare. Mentre, per quanto concerne il giudizio definitivo sull'effetto di alcune scelte particolari, solo l'esperienza maturata sul campo di indagine fornisce delle indicazioni sufficientemente precise ed affidabili. Pertanto, si può pensare che nel predisporre un piano di campionamento sia quanto mai opportuno cercare di fruire

dell'esperienza maturata nel corso di altre rilevazioni statistiche e, in particolare, di non sconvolgere completamente le caratteristiche degli schemi, ai quali si è già fatto ricorso in precedenza.

Data la complessità dell'argomento, non è facile stabilire con esattezza la convenienza di uno schema di campionamento rispetto agli altri e la scelta dipende soprattutto dagli obiettivi che sono ritenuti preminenti. Perciò, una determinata proposta in questo campo probabilmente potrà mettere in evidenza le caratteristiche dello schema preso in esame, ma non potrà mostrare inequivocabilmente la sua convenienza rispetto ad altri schemi di campionamento.

5. Modalità di selezione di un campione areale

La selezione di unità statistiche attraverso campioni casuali semplici è spesso costosa ed in pratica la formazione di un campione nel territorio d'indagine avviene individuando alcuni livelli o stadi di selezione. In ogni stadio o livello si opera mediante la selezione casuale di un determinato numero di unità, ciascuna delle quali è formata da un gruppo di unità appartenenti agli stadi o livelli gerarchicamente inferiori.

Come caso particolare di un campione a due stadi, un campione areale può essere definito individuando un primo stadio di selezione, che spesso viene indicato come lo stadio delle unità primarie di campionamento, le quali sono rappresentate da un determinato numero di aree non troppo grandi di dimensione abbastanza uniforme. All'interno di questo stadio si procede all'estrazione di un certo numero di unità o aree, per poi effettuare una elencazione esaustiva ovvero un microcensimento delle unità di rilevazione appartenenti alle aree estratte. In base ai dati risultanti dal microcensimento è possibile individuare, eventualmente attraverso un'ulteriore operazione di campionamento, il campione finale delle unità di rilevazione.

Per quanto riguarda la formazione di un campione areale con riferimento alle rilevazioni campionarie in Italia, supponendo di far riferimento alle sezioni di censimento come penultimo stadio di selezione delle unità di rilevazione, è necessario definire uno

schema adeguato per la situazione reale e valutare in una certa misura i risultati che si hanno dalla sua applicazione.

Qualora gli scopi della rilevazione fossero quelli connessi con il calcolo delle stime a livello nazionale, l'estrazione casuale delle sezioni di censimento rappresenta una possibilità interessante, ma darebbe luogo ad una dispersione territoriale ampia delle unità di rilevazione. Per ovviare a questo inconveniente sarebbe utile individuare un ulteriore stadio di selezione, quale i comuni, e prevedere come prima fase di selezione l'estrazione di un prestabilito numero di comuni e successivamente la selezione di un determinato numero di sezioni censuarie, come penultimo o secondo stadio di estrazione del campione areale. Infine, per le sezioni incluse nel campione sarebbe opportuno effettuare un microcensimento delle unità di rilevazione, se per un momento ammettiamo che le sezioni di censimento si prestino ad essere sottoposte a tale operazione, per poi procedere all'individuazione delle unità di rilevazione campionarie.

Se anzichè ricavare delle stime nazionali si volessero avere dei risultati affidabili per regione geografica, allora sarebbe necessario predisporre un piano di rilevazione per un campione nazionale stratificato per regione. In questo caso i comuni rappresenterebbero un primo stadio di selezione adeguato per contenere le unità di rilevazione campione in un ambito territoriale sufficientemente ristretto, anche se per diverse regioni (Valle d'Aosta, Friuli V.G., Liguria, Umbria, Molise e Basilicata) l'estrazione dei comuni dovrebbe essere effettuata tra un esiguo numero di unità di campionamento primarie, un numero tanto più esiguo quanto più grande risulta il numero eventuale di strati in cui ripartire i comuni stessi per ridurre la variabilità delle stime afferente all'estrazione dei comuni.

Dalla descrizione sintetica e parziale di questi due schemi di campionamento appare evidente che non si è attribuito un rilievo particolare ad una realtà territoriale come la provincia, che è certamente interessata dalla crescente richiesta di informazioni statistiche.

Un modo di risolvere il problema delle stime a livello delle province è quello di stratificare i comuni, ovvero le unità primarie considerate, per ogni provincia, ma così facendo ci troveremmo, per diverse province, ad estrarre un esiguo numero di comuni all'interno di un altrettanto esiguo numero di comuni. Tale circostanza ha verosimilmente

determinate conseguenze sulla variabilità delle stime per le province più piccole e, pertanto, riteniamo sia opportuno avvalersi di una diversa unità primaria di campionamento.

All'interno delle province, le sezioni di censimento rappresentano un'alternativa ai comuni come unità primarie di campionamento, perché anche se un campione formato a partire dall'estrazione delle sezioni di censimento comporta una certa dispersione delle unità campione nel territorio di indagine, tale dispersione risulta giustificata dalle finalità della rilevazione. Naturalmente, per le province nelle quali il campione ha una piccola dimensione si avranno delle stime non molto precise, ma, qualora se ne ravvisi la necessità, esse possono essere migliorate attraverso un ampliamento del campione. D'altronde, gli effetti di tale ampliamento ricadrebbero anche sulla precisione delle stime regionali e nazionali, dato che queste sarebbero ricavate semplicemente mediante l'aggregazione delle stime provinciali.

Dopo queste considerazioni sommarie su alcuni schemi di campionamento e sulle unità primarie da selezionare, si è ritenuto opportuno riportare una descrizione più dettagliata del problema della selezione delle sezioni di censimento come unità di campionamento primarie all'interno di ciascuna provincia, per avere un'idea dei risultati che ci possiamo attendere.

Circa l'esame della situazione che si presenta in relazione all'uso delle sezioni di censimento come unità primarie, occorre rilevare che non disponiamo di informazioni precise e puntuali sulle caratteristiche della nuova suddivisione territoriale, ma si può fare un'ipotesi sulla situazione nella quale verosimilmente ci si troverà ad operare.

In queste condizioni, le conclusioni alle quali si perviene non sono senz'altro definitive, ma esse possono essere indicative di ciò che si potrebbe fare per cercare di raggiungere gli obiettivi di una rilevazione statistica in alcuni settori di indagine.

Prima di affrontare il problema della formazione del campione dal punto di vista tecnico, notiamo che verosimilmente le sezioni di censimento differiscono l'una dall'altra per molte caratteristiche; perciò la selezione casuale delle sezioni comporta certamente una certa variabilità delle stime campionarie. Praticamente, tale variabilità può essere

controllata (Kish, 1965; Goodman e Kish, 1950) attraverso un'estrazione delle unità con probabilità variabile, oppure ricorrendo al criterio della stratificazione per poi effettuare una selezione casuale all'interno degli strati individuati in base a determinate caratteristiche delle sezioni stesse.

Per quanto riguarda la scelta tra i due accorgimenti menzionati, mediante i quali si può cercare di ridurre la variabilità dei risultati campionari, riteniamo sia preferibile ricorrere alla ripartizione delle sezioni in strati omogenei per diversi motivi. Uno di questi motivi deriva dal fatto che l'estrazione con probabilità variabile è meno semplice dal punto di vista metodologico, mentre un'altra motivazione è di ordine pratico, in quanto si dovrebbe assegnare una probabilità diversa ad ogni sezione, le cui caratteristiche talvolta possono mutare rapidamente e sensibilmente nel tempo.

Volendo fare ricorso alla formazione di un determinato numero di strati, in modo da raggruppare le sezioni con caratteristiche simili, si rende necessario individuare uno o più criteri di stratificazione. La scelta di questi criteri generalmente non è semplice, ma se optiamo per un criterio abbastanza generale, che abbia una certa validità rispetto a diverse variabili oggetto di studio e nei confronti di diverse indagini che si basano sulle stesse unità statistiche di rilevazione, si può pensare di ricorrere alle caratteristiche del comune di appartenenza. Secondo questo criterio, oltre la validità piuttosto ampia, si conseguirebbero anche altri due risultati: di fare riferimento ad una classificazione delle sezioni abbastanza stabile nel tempo (perché la classificazione cambierebbe insieme ai mutamenti che si avrebbero nelle caratteristiche del comune di appartenenza) e di introdurre un accorgimento che comporta una rappresentatività del campione rispetto alle principali tipologie di comuni.

Volendo esemplificare quanto è stato detto, una soluzione operativa è di considerare come criterio di stratificazione delle sezioni, la dimensione del comune di appartenenza in termini del numero di unità di rilevazione, individuando come strato a sé stante le sezioni che appartengono a ciascun comune con popolazione superiore a 20.000 abitanti. Complessivamente, il campione nazionale sarebbe formato estraendo un determinato numero di sezioni da circa 500 strati relativi ai comuni più grandi ed un certo numero di

sezioni dal numero prestabilito di strati, che sono formati da gruppi di sezioni appartenenti ai comuni più piccoli di una medesima provincia.

Lo schema così delineato offre la possibilità di ricavare delle stime anche per i comuni più grandi e di ripartire una certa mole di lavoro sul campo di indagine tra i comuni che (per la raccolta dei dati) posseggono una struttura già collaudata e consolidata nel corso di altre rilevazioni ufficiali.

Se consideriamo come acquisito il modo di procedere descritto sommariamente, vale la pena di prospettare una soluzione in merito all'estrazione casuale delle restanti sezioni di censimento, nelle quali risulta suddiviso il territorio dei comuni più piccoli. Dato che l'obiettivo è di ricavare delle stime a livello di provincia, anche le sezioni dei comuni con meno di 20.000 abitanti potrebbero essere ripartite in strati con l'intento di ridurre la variabilità dovuta alla selezione delle sezioni e per far sì che i comuni con caratteristiche differenti siano adeguatamente rappresentati nel campione.

Pertanto, in ciascuna provincia la stima relativa ad una variabile deriverebbe dalle stime ricavate all'interno di un determinato numero di strati: quelli delle sezioni appartenenti ai comuni più grandi e quelli ottenuti tramite il raggruppamento effettuato per le sezioni appartenenti ai comuni più piccoli.

Operativamente, per la formazione del campione è necessario definire le modalità di selezione delle sezioni di censimento all'interno di ogni strato, mentre per quanto riguarda la raccolta dei dati sul campo i comuni grandi dovrebbero attuare il piano di rilevazione in maniera pressoché autonoma ed i comuni piccoli potrebbero essere guidati e coordinati attraverso un unico organismo (ad esempio la provincia o il comune capoluogo), che eventualmente abbia maturato una certa esperienza quanto mai utile e preziosa nel corso delle operazioni censuarie.

Passando a considerare le caratteristiche delle sezioni di censimento, verosimilmente esse non hanno una dimensione uniforme, sia per i comuni piccoli che per i comuni grandi, mentre il loro numero medio per provincia si potrà aggirare intorno a 4.000. Secondo un'ipotesi abbastanza plausibile, spesso si disporrà per ogni strato di un numero

relativamente elevato di sezioni piccole (in termini del numero di unità di rilevazione) e di un numero variabile di sezioni più grandi.

In queste condizioni, si può ritenere opportuno applicare due procedimenti di selezione differenziati, con l'obbiettivo di pervenire ad un campione autoponderante (per facilitare il calcolo delle stime campionarie) e di determinare gli errori delle stime come una combinazione lineare degli errori riferiti ai vari strati.

La differenza nel procedimento di selezione da applicare per un insieme numeroso di piccole sezioni rispetto a quello da usare per un insieme di grandi sezioni, si giustifica (Kish, 1965) se pensiamo che per le piccole sezioni i dati di sezione siano relativamente poco stabili nel tempo, mentre per le sezioni grandi tali dati sono verosimilmente meno variabili e potrebbero essere usati per una estrazione delle sezioni con probabilità variabile, in modo da ridurre la variabilità dei risultati campionari e controllare la dimensione del campione. Successivamente, nelle sezioni estratte con probabilità variabile sarebbe necessario selezionare un determinato numero (costante) di unità di rilevazione, affinchè si abbia un campione autoponderante. Inoltre, notiamo che per le sezioni piccole l'elencazione delle unità (microcensimento) è una operazione non troppo onerosa, mentre per le sezioni grandi essa risulta generalmente costosa ed è preferibile estrarre direttamente il numero desiderato di unità di rilevazione con le modalità appropriate.

In definitiva, le sezioni piccole non troppo variabili (come dimensione) sarebbero trattate come sezioni di uguale dimensione e la formazione del campione comporterebbe l'estrazione di una frazione costante di unità; ogni sezione estratta (con probabilità uguale) può essere usata per l'aggiornamento dei dati di sezione con le nuove unità, per le quali si dovrebbe procedere ad una estrazione casuale (se il numero di tali unità è piccolo, altrimenti la sezione è da includere in uno strato a parte comprendente i complessi di nuove unità). Invece, le sezioni grandi potrebbero essere estratte con probabilità variabile (proporzionale) con la dimensione, per poi effettuare all'interno delle sezioni estratte la selezione di un numero costante di unità di rilevazione.

Un modo alternativo e pratico di formare il campione per le sezioni grandi è di estrarre in maniera sistematica le unità con un intervallo appropriato di campionamento da

applicare a tutte le sezioni. Così facendo, le unità del campione risulteranno distribuite in tutte le sezioni dello strato considerato.

Passando ad esaminare in dettaglio il problema della scelta casuale delle unità di rilevazione, notiamo che all'interno dello schema descritto è previsto di dover operare sulle unità appartenenti alle sezioni piccole estratte con uguale probabilità, oppure all'interno delle sezioni più grandi. Come si è già detto, le sezioni piccole che risultano selezionate vengono sottoposte ad un microcensimento, dal quale si ricava un elenco delle unità di rilevazione. L'estrazione delle unità da tale elenco può essere attuata considerando le unità singolarmente oppure attraverso dei gruppi o grappoli o segmenti di uguale dimensione g. Qualora si effettui l'estrazione delle unità di rilevazione secondo un criterio sistematico, è sufficiente applicare un intervallo di campionamento di ampiezza K = 1/f, dove f rappresenta la frazione di campionamento. Se invece si preferisce estrarre dei grappoli o segmenti di unità, per ciascuna sezione campione è sufficiente dividere l'elenco delle unità in segmenti per poi applicare la stessa frazione f di campionamento ai segmenti nei quali risulta suddivisa la sezione stessa.

Per l'estrazione delle unità all'interno delle sezioni grandi, se supponiamo di effettuare una selezione sistematica delle unità campionarie da tutte le sezioni, l'estrazione delle unità può essere effettuata vantaggiosamente dal punto di vista finanziario, estraendo le unità di rilevazione a gruppi di g unità, mediante un intervallo di campionamento di ampiezza g/f (Kish, 1965). In questo caso, ad ogni unità estratta si farebbe corrispondere un gruppo o segmento di g unità, che risultano individuate considerando le unità comprese tra le unità che occupano i posti rg/f e rg/f+g (r =1,2, ..., F_S f/g; con F_S pari al numero di unità nello strato s), come se si avesse di fronte un segmento aperto a destra. La tecnica descritta non presenta grandi difficoltà, purchè sia possibile stabilire un criterio meccanico per ordinare le unità di rilevazione secondo una prefissata direzione e per attribuire le unità nuove o mancanti in maniera univoca in base a dei criteri prefissati. In altre parole si provvederebbe all'individuazione del numero desiderato di segmenti, che faciliterebbe la raccolta dei dati e che sarebbero utili anche per l'aggiornamento del campione con le unità nuove o mancanti (di nuovo, se il numero di queste unità è limitato,

altrimenti tali segmenti dovrebbero essere inclusi nello strato a parte dei complessi di nuove unità).

Naturalmente, l'uso dei segmenti di unità comporta generalmente un incremento della variabilità delle stime campionarie, rispetto a quella che caratterizza l'estrazione di singole unità, ma permette di conseguire alcuni vantaggi, tra i quali quello di rendere più agevoli le operazioni da svolgere per la raccolta dei dati.

In ogni caso, la base di campionamento necessaria è rappresentata dai dati dell'ultimo censimento, che può anche essere sottoposto ad un controllo mediante la rilevazione campionaria, qualora questa venga effettuata in tempi ravvicinati a quello di riferimento per le operazioni censuarie.

Come abbiamo detto in precedenza, lo schema previsto di campionamento di tipo areale si avvale di un certo numero di segmenti di dimensione g. Circa i vantaggi e gli svantaggi derivanti dalla selezione di tali segmenti si possono fare diverse considerazioni, che riguardano la praticità della soluzione adottata e l'effetto che si riscontra in relazione alla variabilità delle stime. Infatti, come è facile da prevedere, l'individuazione delle unità di rilevazione è più semplice e rapida, dà luogo raramente a sostituzioni e, inoltre, l'uso dei segmenti consente di aggiornare i dati con le unità nuove o mancanti, seguendo delle regole prefissate per la loro attribuzione ai vari segmenti inseriti nel campione. Dal punto di vista dei risultati campionari il ricorso ai segmenti introduce una fonte di variabilità, che può influire sull'efficienza in senso statistico degli stimatori per alcune variabili. Tuttavia, non va dimenticato che il riferimento a segmenti e sezioni consente di ovviare in una certa misura ai problemi di invecchiamento nel tempo dei registri delle unità di rilevazione, anche se ciò necessita di un aggiornamento periodico (ad esempio annuale) dei dati, al fine di mantenere delle liste aggiornate delle unità appartenenti ai segmenti ed alle sezioni che risultano inseriti nel campione.

Nel delineare alcuni tratti essenziali di un campione areale basato sulle sezioni di censimento, come unità di campionamento primarie, non abbiamo accennato a diversi aspetti pratici di natura particolare, quali la necessità di riferire ogni unità ad una collocazione territoriale unica, l'opportunità di predisporre un elenco delle localizzazioni

potenzialmente interessate dalla rilevazione ed il trattamento dei complessi di unità nuove o mancanti, rispetto a quelle che risultano inserite nella base di campionamento. Per una trattazione adeguata di tali problemi riteniamo non si possa prescindere da una sperimentazione diretta, ma ci sembra che, nel complesso, un campione areale definito sulla base delle sezioni di censimento meriti una certa attenzione, qualora sia in grado di fornire delle informazioni affidabili e sufficientemente precise per diverse esigenze di documentazione statistica. Inoltre, vale la pena di notare che la qualità di tali informazioni migliorerebbe col tempo, mano a mano che l'esperienza suggerisce particolari accorgimenti e che la definizione delle sezioni diventa più appropriata e si conforma alle indicazioni fornite dall'Istat.

6. Uno schema teorico di campione areale

Lo schema di campionamento descritto informalmente per ogni provincia si basa sulle sezioni di censimento come unità primarie di campionamento e sulla loro suddivisione in segmenti uguali. L'applicazione di tale schema può essere formalizzata introducendo alcune ipotesi semplificatrici della situazione reale nella quale dobbiamo operare. Naturalmente, in pratica occorrerà studiare l'effetto di tali ipotesi attraverso l'esperienza empirica, ma a priori è possibile fare una valutazione approssimata dei risultati che ci possiamo attendere.

Allo scopo di affrontare il problema della stima e della valutazione dell'errore campionario, desideriamo trattare il caso concreto di un'indagine che ha come unità di rilevazione la famiglia e facciamo riferimento alla seguente notazione:

numero di famiglie nella provincia	F
numero di strati delle sezioni	S
numero di famiglie dello strato s	F_s
numero delle famiglie della sezione j dello strato s	F_{sj}
numero di famiglie di ogni segmento	g
numero medio di famiglie per sezione dello strato s	$\overline{F}_{s.}$
numero di individui nella provincia	N
numero di individui nello strato s	N_s
numero di individui della sezione j, strato s	N_{sj}

numero di individui del segmento k, sezione j, strato s	N_{sjk}
numero di individui della famiglia l , segmento k, sezione j, strato s	N_{sjkl}
frazione di campionamento delle famiglie	f
frazione di campionamento delle sezioni	f_1
frazione di campionamento all'interno delle sezioni	f_2
numero di sezioni della provincia	A
numero di sezioni dello strato s (s=1, 2,..., S)	$A_s = F_s/\overline{F}_{s.}$
numero di segmenti della sezione j dello strato s	A_{sj}
numero medio di individui per sezione dello strato s	$\overline{N}_{s.} = N_s/A_s$
numero medio di individui per segmento, sezione j, strato s	$\overline{N}_{sj.} = N_{sj}/A_{sj}$

Come è noto, diverse rilevazioni campionarie dell'Istat fanno uso della famiglia come unità di rilevazione nell'ambito di schemi di campionamento diversi da quello areale. Per diversi campi di indagine, tali rilevazioni forniscono delle informazioni ritenute generalmente affidabili, perciò il caso preso in esame rappresenta un interessante termine di raffronto per cercare di capire se lo schema di campionamento areale può essere considerato uno strumento alternativo da impiegare nello svolgimento di alcune attività di documentazione statistica.

Passando a considerare il problema delle stime campionarie, due espressioni alle quali è utile fare riferimento sono quella della stima

$$n_{sj} = \frac{1}{f_2} \sum_{k=1}^{A_{sj}f_2} N_{sjk} \qquad (4)$$

del numero N_{sj} di individui della sezione j appartenente allo strato s e quella della stima

$$n_s = \frac{1}{f_1} \sum_{j=1}^{A_s f_1} n_{sj} = \frac{1}{f_1}\frac{1}{f_2} \sum_{j=1}^{A_s f_1} \sum_{k=1}^{A_{sj}f_2} N_{sjk} \qquad (5)$$

del numero N_s di individui dello strato s, dove $1/f_1$ e $1/f_2$ rappresentano, rispettivamente, gli intervalli di campionamento nel primo stadio (sezioni) e nel secondo e ultimo stadio (segmenti) di campionamento.

La varianza dello stimatore n_s è fornita dall'espressione seguente (nel caso di estrazione senza ripetizione)

$$\text{var}(n_s) = \frac{A_s - A_s f_1}{A_s - 1} \frac{A_s}{f_1} \sigma^2_{sj} + \frac{A_s}{a_s} \sum_{j=1}^{A_s} \frac{A^2_{sj}}{a_{sj}} \frac{A_{sj} - A_{sj} f_2}{A_{sj} - 1} \sigma^2_{sjk} \qquad (6)$$

dove i simboli σ^2_{sj} , σ^2_{sjk}

$$\sigma^2_{sj} = \frac{1}{A_s} \sum_{j=1}^{A_s} (N_{sj} - \overline{N}_{s.})^2 \qquad (7)$$

$$\sigma^2_{sjk} = \frac{1}{A_{sj}} \sum_{k=1}^{A_{sj}} (N_{sjk} - \overline{N}_{sj.})^2 \qquad (8)$$

rappresentano la varianza del numero di individui, rispettivamente, tra sezioni nello strato s e tra segmenti nella sezione j.

Circa la stima n del numero totale N degli individui di una provincia, essa

$$n = \sum_{s=1}^{S} n_s \qquad (9)$$

risulta dalla somma delle stime n_s calcolata per tutti gli strati, mentre la sua varianza

$$\text{var}(n) = \sum_{s=1}^{S} \text{var}(n_s) \qquad (10)$$

è ottenuta semplicemente come somma delle varianze delle stime n_s in tutti gli S strati.

Per quanto riguarda le quantità σ^2_{sj} , σ^2_{sjk}, notiamo che, indicando con $\overline{A}_{s.}$ il numero medio di segmenti per sezione, esse possono essere scritte diversamente

$$\sigma^2_{sj} = \text{var}(N_{sj}) = \text{var}(\sum_{k=1}^{A_{sj}} N_{sjk}) = \overline{A}_{s.}\ g\ \text{var}(N_{sjkl}) \qquad (11)$$

$$\sigma^2_{sjk} = \text{var}\,(N_{sjk}) = g\ \text{var}\,(N_{sjkl}) \qquad (12)$$

se ammettiamo che sia nulla o trascurabile la correlazione tra i numeri N_{sjk} e $N_{sjk'}$ di individui di due segmenti k e k' ($k \neq k'$) e quella tra i numeri N_{sjkl} e $N_{sjkl'}$ di individui delle due famiglie l, l' ($l \neq l'$).

Mediante le espressioni (11,12) la varianza dello stimatore n_s (omettendo i fattori di correzione dovuti all'estrazione senza ripetizione)

$$\text{var}\,(n_s) = A_s \frac{1}{f_1} \overline{A}_s.\ g\ \text{var}(N_{sjkl}) + A_s\, \overline{A}_s. \frac{1}{f_1} \frac{1}{f_2} g\ \text{var}(N_{sjkl}) \qquad (13)$$

risulta nuovamente dalla somma di due componenti, la prima delle quali deriva dalla variabilità tra le sezioni, mentre la seconda è dovuta alla variabilità tra segmenti all'interno delle sezioni. In questa forma la varianza dello stimatore è funzione della varianza della distribuzione delle famiglie secondo il numero di individui N_{sjkl} e sarà tanto più grande rispetto a quella che caratterizza un campione casuale semplice, quanto più grande è la componente della variabilità dovuta all'estrazione casuale delle sezioni.

Nella descrizione formale di alcuni dei problemi inerenti gli stimatori ed i relativi errori campionari non abbiamo preso in esame diversi aspetti importanti per l'applicazione dello schema di campionamento previsto in questo contesto. In particolare, affinchè la trattazione dell'argomento sia operativamente utile, per ciascuno strato è necessario stabilire il modo di ripartire le unità da inserire nel campione tra le sezioni ed i segmenti, ovvero di fissare le frazioni di campionamento delle sezioni e delle unità.

Supposto di aver fissato per la provincia considerata una frazione f complessiva di campionamento pari alla frazione nazionale di campionamento, se facciamo riferimento al caso di due stadi di selezione del campione, la frazione f

$$f = f_1\, f_2 \qquad (14)$$

risulta uguale al prodotto tra la frazione f_1 di campionamento nel primo stadio (le sezioni) e la frazione f_2 di campionamento nel secondo stadio delle unità di rilevazione. Qualora le unità siano estratte in maniera sistematica a gruppi di g unità, la frazione f può essere riscritta

$$f = f_1 \, g \, f_g \qquad (15)$$

considerando che la frazione f_2 risulta pari ad un multiplo della frazione f_g delle unità da estrarre, dato che ad ogni unità selezionata corrisponde un segmento di g unità di rilevazione. Tali frazioni possono essere usate per le operazioni di campionamento in ciascuno strato di sezioni all'interno di una determinata provincia.

I costi da sostenere per la raccolta di dati e la variabilità che caratterizza gli stimatori all'interno di un determinato schema rappresentano due elementi importanti per stabilire il numero di sezioni che è opportuno selezionare e per assicurare una distribuzione territoriale del campione.

Per ciò che concerne un campione casuale semplice, notiamo che le $F_s f$ unità di rilevazione da estrarre tra tutte le F_s unità nello strato s risultano potenzialmente distribuite in altrettante sezioni delle A_s sezioni di censimento (se $A_s \geq F_s f$), mentre se estraessimo casualmente una frazione f_1 di sezioni e successivamente una frazione f_2 di unità dalle sezioni estratte, allora le $F_s f$ unità campione apparterrebbero ad un numero a_s di sezioni ($a_s = = f_1 A_s$). Il costo per la raccolta dei dati è ovviamente differente per i due tipi di campione, qualora si pensi che esso sia scomponibile in due parti, ovvero il costo C_a per contattare una sezione e il costo C_u per rilevare ciascuna unità. Naturalmente, la valutazione di tali costi è inevitabilmente approssimativa, in quanto, ad esempio, la spesa C_a riguarda il tragitto da compiere e il tempo richiesto per raggiungere la sezione di censimento tutte le volte che ciò si rende necessario, mentre la spesa C_u riguarda essenzialmente il tempo impiegato per reperire i dati presso l'unità di rilevazione quando è già stata localizzata. Inoltre, non va dimenticato che le spese indicate non includono quelle relative alla gestione dello schema di campionamento, anche se queste ultime hanno verosimilmente un'incidenza marginale sul costo complessivo di un campione abbastanza ampio. Pertanto, con una certa cautela, si può pensare che la differenza

$$H = F_s f \, (C_a + C_u) - (a_s C_a + F_s f \, C_u) \qquad (16)$$

tra i costi relativi ai due campioni sia funzione dei costi unitari C_a, C_u e, in termini finanziari, essa esprime la convenienza di raccogliere i dati in a_s sezioni anzichè in un numero $F_s f$ di sezioni di censimento.

Oltre l'aspetto finanziario, un altro elemento da prendere in esame è la variabilità che caratterizza gli stimatori, perché l'estrazione casuale delle sezioni di censimento produce generalmente un aumento della variabilità che si avrebbe usando un campione casuale

semplice, cosicché per esaminare l'efficienza in senso lato delle stime occorre confrontare il prodotto

$$\sigma_c F_s f (C_a + C_u) \tag{17}$$

tra errore (σ_c) e costo di raccolta dei dati per un campione casuale semplice di unità appartenenti a $F_s f$ sezioni, col prodotto

$$\sigma_g (a_s C_a + F_s f C_u) \tag{18}$$

tra errore (σ_g) e costo di raccolta dei dati attraverso un campione di $F_s f / a_s$ unità in ciascuna delle a_s sezioni. Il rapporto Q

$$Q = \frac{\sigma_c}{\sigma_g} \frac{F_s f (C_a + C_u)}{(a_s C_a + F_s f C_u)} \tag{19}$$

tra le espressioni (17) e (18) fornisce un'indicazione sulla convenienza derivante dall'uso di un campione casuale semplice ($Q < 1$) o di un campione a due stadi ($Q > 1$).

Come si vede facilmente, tale rapporto consente di stabilire un limite per il numero di sezioni

$$a_s < F_s f \left(\frac{\sigma_c}{\sigma_g} \frac{C_a + C_u}{C_a} - \frac{C_u}{C_a} \right) \tag{20}$$

che si devono estrarre affinchè il campione ottenuto attraverso a_s sezioni sia più conveniente del campione casuale semplice di unità all'interno di un certo strato di sezioni.

In queste condizioni, ad esempio, se l'effetto del disegno di campionamento

$$D^2 = \frac{\sigma_g^2}{\sigma_c^2} \tag{21}$$

relativo al confronto tra l'errore σ_g di uno schema particolare di campionamento e l'errore σ_c di un campione casuale semplice è pari a 4, mentre il rapporto tra il costo C_u per

rilevare un'unità ed il costo C_a per contattare una sezione è pari a 1/5, allora si rileva che il numero a_s

$$a_s < F_s f\, 2 / 5 \tag{22}$$

di sezioni da estrarre deve essere inferiore a 2/5 della dimensione del campione. Pertanto, l'espressione (20) fornisce un limite superiore al numero di sezioni campione di cui è opportuno avvalersi, in quanto al di sopra di tale limite il risparmio in termini finanziari non compensa la perdita, che, in termini di variabilità delle stime, deriva dalla rinuncia ad usare un campione casuale semplice.

La descrizione riportata in queste pagine si riferisce ad una prima parte di uno schema di campionamento di unità di rilevazione appartenenti ad un insieme numeroso di sezioni piccole e di dimensione uniforme. L'estrazione delle sezioni viene effettuata con uguale probabilità, ma non è stata fornita un'indicazione precisa sul modo di determinare il numero di sezioni e, all'interno delle sezioni estratte, il numero di unità di rilevazione da selezionare, qualora sia stata fissata la frazione f complessiva di campionamento.

Questo problema può essere affrontato (Kish, 1965; Hansen et al., 1953) calcolando il numero F_0 ottimo di unità di rilevazione da estrarre per ogni sezione. Come si è già osservato, si può supporre che il costo complessivo

$$C = a_s C_a + a_s F_0 C_u \tag{23}$$

per la raccolta dei dati sul campo dipenda dal numero di sezioni a_s, dal numero $F_s f = a_s F_0$ di unità da rilevare a gruppi di F_0 unità e dai costi C_a e C_u inerenti, rispettivamente, al lavoro per contattare una sezione e a quello per rilevare un'unità.

Inoltre, se supponiamo che le sezioni abbiano la stessa dimensione, possiamo porre $A_{sj} = \overline{F}_{s.}$ e $a_{sj} = F_0$ (j = 1, 2, ..., A_s). Allora, un'espressione della varianza

$$\mathrm{var}(n_s) = \frac{A_s - a_s}{A_s - 1} \frac{A_s^2}{a_s} \sigma_{sj}^2 + \frac{\overline{F}_{s.} - F_0}{\overline{F}_{s.} - 1} \frac{1}{a_s F_0} A_s \overline{F}_{s.}^2 \sum_{j=1}^{A_s} \sigma_{sjk}^2 \tag{24}$$

dello stimatore n_s del numero N_s di individui è data dalla somma di due termini che rappresentano la variabilità afferente all'estrazione, rispettivamente, nel primo e nel secondo stadio di selezione delle unità di rilevazione, come se si fosse operato con dei segmenti di dimensione pari ad uno.

In tali condizioni, il numero F_0 ottimo di unità da estrarre per ogni sezione è dato da quel valore che, fissato il costo C, rende minimo il valore dell'espressione (24) della varianza dello stimatore n_s (oppure che, stabilito un valore della varianza, rende minimo il costo C).

Analiticamente, al variare di a_s e F_o la soluzione del problema è ottenuta determinando il valore minimo dell'espressione seguente,

$$E = \text{var}\,(n_s) + \lambda\,(a_s C_a + a_s F_o C_u - C) \qquad (25)$$

dove il simbolo λ rappresenta il moltiplicatore di Lagrange applicato all'equazione inerente al vincolo imposto sui costi di rilevazione.

Secondo tale procedimento si ricava il numero F^*

$$F^* = \left(\frac{\sigma_e^2}{\sigma_a^2 - \frac{\sigma_e^2}{\overline{F}_{s.}}} \frac{C_a}{C_u} \right)^{1/2} = \left(\frac{1-\rho}{\rho} \frac{C_a}{C_u} \right)^{1/2} \qquad (26)$$

più conveniente di unità da estrarre in ciascuna delle sezioni campione, che dipende dalla variabilità interna (σ_e^2) alle sezioni, dalla variabilità tra sezioni (σ_a^2), dal numero medio $\overline{F}_{s.}$ di unità per sezione e dal rapporto tra i costi C_a e C_u ; inoltre, notiamo che il numero F^* di unità campione per sezione può essere scritto in funzione del coefficiente ρ di correlazione intraclasse, che rappresenta una misura dell'omogeneità interna alle sezioni e che risulta differente a seconda della variabile oggetto di studio.

Naturalmente, per il calcolo del valore F^* occorrerebbe tenere conto dei valori assunti da ρ per diverse variabili e della situazione in cui si opera, che mediamente supponiamo sia caratterizzata da uno dei comuni grandi, formato da uno strato di sezioni con F = 20.000 famiglie distribuite in circa 400 sezioni, ciascuna delle quali comprende mediamente 50 famiglie.

Se prendiamo come riferimento il caso $\rho = 0{,}10$, $C_u/C_a = 1/5$, allora si ha $F^* \cong 7$ e una soluzione approssimata è di porre il numero F_0 di unità di rilevazione pari a 10

$$F_o = 10 \qquad (27)$$

ed il numero a_s di sezioni campione pari a 14,

$$a_S = \frac{Ff}{F_0} = 14 \qquad (28)$$

che si ottiene in base ad una frazione complessiva f = 0,007 di campionamento.

La dimensione $F_0 = 10$ può essere ritenuta conveniente considerando che in pratica ogni segmento potrà avere una dimensione effettiva leggermente superiore o inferiore a quella richiesta, a causa della presenza di qualche unità nuova o mancante e/o dell'inesistenza di alcune unità. Inoltre, essa è tale da rendere economicamente meno onerosa la raccolta dei dati, in quanto per ogni spostamento un rilevatore è in grado di svolgere il lavoro connesso alla rilevazione delle dieci unità appartenenti ad un determinato segmento (ad esempio consegna e ritiro dei questionari di indagine).

Infine, rileviamo che se si fosse calcolato F_0 in funzione di un valore $\rho = 0,3$ si sarebbe ottenuta una dimensione dei segmenti di circa 3 unità, ossia una dimensione troppo piccola per fruire della convenienza economica dovuta al fatto di far riferimento a dei segmenti per la raccolta dei dati.

Con i valori indicati per il numero F_0 di unità e per il numero a_s di sezioni, il rapporto

$$Q' = \frac{F_S f\,(C_a + C_u)}{a_S C_a + F_S f\, C_u} = 4 \qquad (29)$$

tra i costi di raccolta dei dati nei due casi è pari a quattro, mentre il rapporto

$$Q'' = \frac{\sigma_c}{\sigma_g} = \frac{1}{D} \cong \frac{1}{2} \qquad (30)$$

tra gli errori delle stime, ovvero la radice quadrata del reciproco dell'effetto

$$D^2 = 1 + \rho\,(F_0 - 1) \qquad (31)$$

dovuto allo schema di campionamento, risulta approssimativamente pari a 1/2 qualora si faccia riferimento al caso $\rho = 0,3$. Infine, il rapporto tra i prodotti costo-errore

$$Q = \frac{\sigma_c}{\sigma_g}\,\frac{F_S f\,(C_a + C_u)}{a_S\, C_a + F_S f\, C_u} \cong 2 \qquad (32)$$

è di circa due unità e denota l'esistenza di un vantaggio (Q > 1) derivante dall'uso di due stadi di selezione nel campionamento delle unità di rilevazione.

Notiamo che secondo le esperienze effettuate (Kish, 1987; Kish e Frankel, 1970) con questo tipo di campioni, quasi tutte le variabili danno luogo ad un valore del coefficiente ρ inferiore a 0,3; pertanto, l'effetto D^2 dello schema di campionamento risulta più piccolo, il rapporto Q" più elevato ed il valore di Q sarà generalmente anche più grande di quello indicato con l'espressione (32), cosicché l'uso del campione a due stadi risulta spesso più conveniente di quanto appare in base ai calcoli riportati.

7. Schema di campionamento per i comuni non autorappresentativi

Il procedimento di estrazione che è stato descritto può essere adeguato per la formazione di un campione di unità di rilevazione appartenenti a comuni di una certa dimensione e con un numero abbastanza ampio di sezioni aventi una dimensione non troppo diversa.

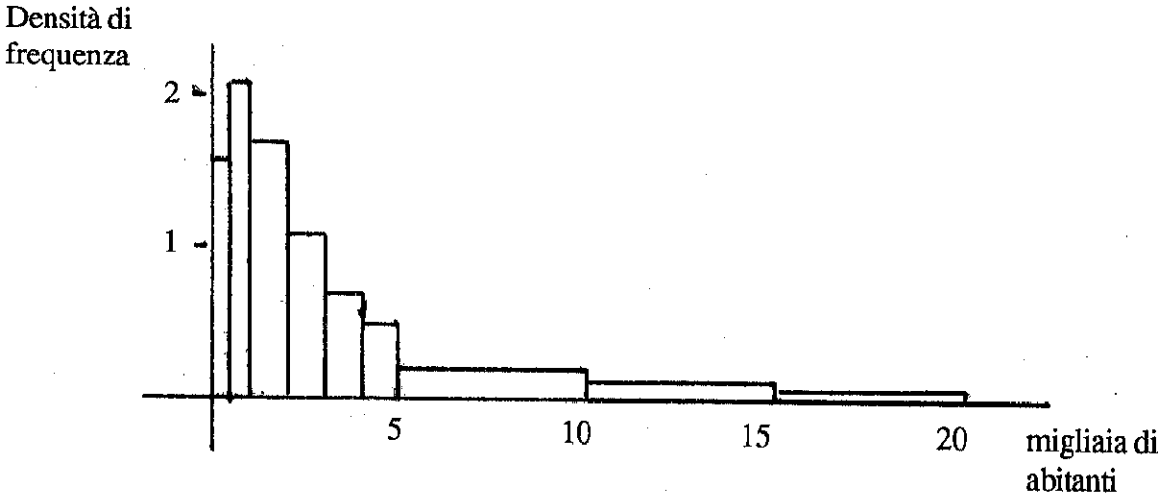
Se passiamo a considerare uno schema di campionamento da utilizzare all'interno di ogni provincia per i piccoli comuni, la prima idea che viene in mente è procedere ad una estrazione casuale di tali comuni con probabilità variabile o costante, per poi effettuare la selezione delle unità di rilevazione all'interno dei comuni estratti seguendo le modalità opportune per predisporre un campione efficiente. Perciò, prima di proseguire su questa via, è certamente interessante fare qualche riflessione sulle caratteristiche dei comuni con meno di 20.000 abitanti.

Dalla distribuzione dei comuni piccoli (con meno di 20.000 abitanti) secondo il numero di abitanti al 31 dicembre dell'anno 1989 in Italia (fig.1), si nota una accentuata asimmetria positiva e si ha l'impressione che un'estrazione casuale semplice dei comuni darebbe luogo ad una certa variabilità dei risultati campionari. Volendo cercare di ridurre questa fonte di variabilità, un modo di operare è quello di individuare alcuni gruppi omogenei di comuni rispetto alla loro dimensione. Per il complesso di tutti i comuni, ad esempio, di potrebbe pensare di formare due strati di comuni, il primo dei quali comprende circa 5.900 comuni (circa 300 per regione ovvero 60 per provincia) con popolazione inferiore a 5.000 abitanti (per un totale di circa 11 milioni di abitanti), mentre il secondo gruppo o strato risulterebbe formato da circa 1.700 comuni (circa 85 per regione, ovvero 17 per provincia) con un numero di abitanti compreso tra 5.000 e 20.000

Tab.1 - Distribuzione dei comuni con un numero di abitanti non superiore a 20.000, secondo la classe di ampiezza demografica (Italia, 31-12-1989)

Classi	Frequenza
Fino a 500	799
501-1000	1.143
1001-2000	1.715
2001-3000	1.069
3001-4000	660
4001-5000	528
5001-10000	1.149
10001-15000	398
15001-20000	190
Totale	7.651

Fig.1 - Distribuzione dei comuni con un numero di abitanti non superiore a 20.000 secondo la classe di ampiezza demografica (Italia, 31-12-1989)



Fonte dei dati: Istat, *Compendio Statistico Italiano*, ed.1991

(per un totale di circa 16 milioni di abitanti). In questa situazione, la formazione degli strati permetterebbe di controllare la variabilità dei risultati campionari, che è dovuta all'estrazione casuale dei comuni. Mentre la scelta della frazione di campionamento potrebbe essere fatta in base all'importanza degli strati in termini dell'ammontare complessivo della popolazione. Inoltre, riteniamo che all'interno dei comuni estratti sarebbe opportuno procedere all'estrazione di un numero adeguato di unità di rilevazione, in modo da mantenere costante la probabilità di inserimento nel campione per tutte le unità appartenenti alla popolazione statistica.

Come è facile intuire, i comuni estratti possono essere confinanti oppure distanti, cosicché la dispersione territoriale delle unità campione può risultare più o meno ampia per effetto del caso; inoltre notiamo che, operando a livello di regione o di provincia, ci troveremmo di fronte a situazioni particolari, che andrebbero esaminate caso per caso, soprattutto in funzione del numero e della dimensione dei comuni. Perciò, riteniamo sia utile individuare uno schema diverso di estrazione del campione e applicarlo in maniera uniforme alle province italiane.

Per fissare alcuni punti di riferimento, notiamo che mediamente ogni provincia è formata da circa 5 comuni autorappresentativi (se in totale per l'Italia consideriamo 500 comuni con più di 20.000 abitanti su 100 province) e da circa 75 piccoli comuni (circa 7.500 comuni con meno di 20.000 abitanti su 100 province). Se consideriamo una provincia con simili caratteristiche come una situazione tipica, possiamo predisporre un campione al fine di ottenere delle stime, che risultano aggregabili qualora si vogliano effettuare delle stime regionali e nazionali.

Per quanto concerne i comuni autorappresentativi, il campione può essere ricavato seguendo le linee tratteggiate in precedenza e spesso le stime a livello di comune presenteranno una certa variabilità, in quanto la dimensione del campione varia a seconda della provincia e del comune in funzione dell'ammontare della popolazione. Invece, per la scelta casuale delle unità di rilevazione appartenenti ai piccoli comuni di una provincia, ci sembra che sostanzialmente vi siano due soluzioni per ridurre i costi di raccolta dei dati e la variabilità delle stime. Un modo di procedere consiste nell'estrazione casuale con probabilità variabile di un certo numero di comuni e successivamente nella selezione di un certo numero di unità di rilevazione tra quelle appartenenti ai comuni estratti. Tuttavia, oltre gli inconvenienti già menzionati in relazione alla selezione dei comuni, notiamo che nel primo stadio di campionamento si dovrebbe effettuare un'estrazione tra un numero limitato di unità (comuni) e che si incontrerebbero alcuni problemi pratici, volendo

ricorrere ad un campione autoponderante o alla formazione di strati omogenei delle unità di primo stadio.

Un altro modo di selezionare le unità di rilevazione appartenenti ai piccoli comuni è quello di utilizzare, come primo stadio per la formazione del campione, le sezioni di censimento, che potrebbero essere stratificate in base alla loro dimensione e/o alla dimensione dei comuni di appartenenza.

In ambedue i casi è facile prevedere una certa variabilità delle stime a livello di provincia, ma sussiste una differenza evidente tra i due procedimenti di estrazione per quanto riguarda il numero dei comuni interessati dalla rilevazione campionaria, in quanto nel primo caso tale numero è stabilito a priori, mentre esso risulta determinato a posteriori, qualora si considerino le sezioni di censimento come primo stadio di selezione delle unità di rilevazione.

Nel complesso, per i comuni non autorappresentativi ci sembra opportuno fare riferimento nell'ambito di ogni provincia alle sezioni di censimento come primo stadio di estrazione del campione, perché esso presenta un certo margine di scelta per ricorrere al criterio della stratificazione e consente di ottenere una maggiore distribuzione territoriale delle unità di rilevazione inserite nel campione.

Volendo dare un'idea più circonstanziata dello schema di campionamento areale e dei risultati conseguibili nell'ambito di una rilevazione sulla popolazione, si possono fare alcune ipotesi sulle condizioni in cui praticamente si dovrà operare.

Come punti di riferimento per fare alcune considerazioni sulle stime e sugli errori, notiamo che circa 27 milioni di abitanti risultano distribuiti in circa 7.500 comuni piccoli. Di tali abitanti, circa 2/5 (11 milioni) si trovano nei comuni (circa 5.900) con meno di 5.000 abitanti, mentre i restanti 3/5 (16 milioni) si trovano nei comuni (circa 1.700) con una popolazione compresa tra 5.000 e 20.000 abitanti. Pertanto, per ogni provincia mediamente si può pensare che vi sia un numero di 59 comuni con meno di 5.000 abitanti (in totale circa 110.000) ed un numero di 17 comuni con un numero di abitanti compreso tra 5.000 e 20.000 (in totale circa 160.000 abitanti).

Per quanto riguarda la popolazione delle sezioni di censimento, se facciamo riferimento a 400.000 sezioni con una popolazione media di 150 abitanti (50 famiglie), è abbastanza agevole effettuare alcuni calcoli approssimativi, per cercare di intravedere le caratteristiche dei risultati attesi dall'applicazione dello schema descritto o da uno schema simile, qualora si dovessero introdurre degli accorgimenti, soprattutto per tenere conto delle diversità esistenti tra le dimensioni delle sezioni in termini dell'ammontare della

popolazione, ad esempio mediante una ripartizione delle sezioni in alcuni strati o gruppi omogenei.

Naturalmente, attraverso gli elementi forniti al riguardo di un eventuale schema di campionamento areale sulla popolazione non si hanno delle informazioni precise sui risultati ai quali si perverrebbe mediante la sua applicazione, in quanto per un esame circostanziato del problema è necessaria una sperimentazione per approfondire lo studio dell'effetto prodotto sui risultati da diverse condizioni o scelte particolari, quali la dimensione del campione di sezioni, il numero di unità di rilevazione per sezione e, in generale, delle modalità di estrazione delle unità di rilevazione. Tuttavia, è senz'altro utile disporre di una descrizione sommaria e approssimativa sulle caratteristiche della rilevazione statistica prospettata.

Circa la dimensione del campione, facciamo riferimento ad una frazione di campionamento pari a 0,007 e quindi ad un campione di circa 189.000 abitanti su una popolazione di 27 milioni di abitanti nei comuni di piccola e media ampiezza. Tale dimensione corrisponde ad un campione di circa 63.000 famiglie, delle quali una frazione pari a circa 2/5 (11 su 27 milioni) devono essere estratte tra le famiglie appartenenti alle sezioni dei comuni con meno di 5.000 abitanti, mentre per circa 3/5 esse dovranno appartenere ai comuni con un numero di abitanti compreso tra 5.000 e 20.000. Ciò, corrisponde ad un campione per provincia di 770 abitanti (circa 260 famiglie) da estrarre nello strato delle sezioni (110.000/150 ≅ 740 sezioni) appartenenti ai comuni più piccoli con una popolazione media per provincia di 110.000 abitanti, mentre in ogni provincia per lo strato delle sezioni (160.000/150 ≅ 1060) appartenenti ai comuni di dimensione intermedia si avrebbe un campione di 1.120 abitanti (circa 370 famiglie) su 160.000 abitanti.

Come è facile vedere, si è applicato un criterio di stratificazione proporzionale, ma lo schema è suscettibile di alcuni miglioramenti, qualora si disponga di informazioni affidabili sulla variabilità che caratterizza i due strati. In ogni caso, il passo successivo consiste nello stabilire il numero di sezioni campione tra le quali ripartire per ciascuno strato, rispettivamente, le 260 e 370 famiglie campione dei comuni di piccola e media ampiezza della provincia considerata, al fine di ottenere un campione efficiente in senso lato (costo-errore).

Secondo quanto è stato fatto in precedenza per la formazione del campione nei comuni con oltre 20.000 abitanti, in ogni strato di sezioni appartenenti ai piccoli comuni il numero di sezioni da estrarre e, all'interno di queste, il numero di unità di rilevazione da

numero di sezioni da estrarre e, all'interno di queste, il numero di unità di rilevazione da inserire nel campione possono essere calcolati in funzione dei costi per la raccolta dei dati e della varianza che caratterizza le stime campionarie.

Tenuto conto del fatto che l'estrazione delle sezioni in un prefissato strato di sezioni interessa comuni diversi, supponiamo che il rapporto tra i costi C_{a1} e C_u, rispettivamente per contattare una sezione e per raccogliere i dati presso un'unità di rilevazione, sia di 10 a 1; inoltre, facciamo riferimento ad un valore pari a 0,10 del coefficiente ρ di correlazione intraclasse, che in molti casi può essere ritenuto non troppo diverso da quello reale (Kish, 1987). In queste condizioni, il numero conveniente

$$F^{**} = \sqrt{\frac{C_{a1}}{C_u}\frac{1-\rho}{\rho}} \cong 10 \qquad (33)$$

di unità da estrarre in ogni sezione campione risulta approssimativamente pari a 10 e risulterebbe tanto più elevato quanto più piccolo è il valore di ρ. Tuttavia, riteniamo che il valore $F_{o1}=10$ sia adeguato per contenere il costo di rilevazione

$$C = \frac{F_s f}{F_{o1}} C_{a1} + F_s f\, C_u \qquad (34)$$

e l'effetto

$$D^2 = \frac{\sigma^2_{a1}}{\sigma^2_c} = 1 + \rho\,(F_{o1} - 1) \qquad (35)$$

dovuto all'uso di uno schema particolare a1 di campionamento rispetto allo schema c, che è rappresentato da un campione casuale semplice.

Sulla base di questi elementi il rapporto costo-errore

$$Q = \frac{\sigma_c}{\sigma_{a1}} \frac{F_s f\,(C_{a1} + C_u)}{a_s C_{a1} + F_s f\, C_u} \cong 3 \qquad (36)$$

tra i due schemi a1 e c di campionamento risulta approssimativamente pari a 3 per ambedue gli strati di sezioni presi in esame e dipende dalla valutazione effettuata per i costi C_{a1} e C_u, oltre che dall'effetto dello schema di campionamento sulla variabilità delle

conveniente (Q > 1), anche se una valutazione precisa del rapporto Q può essere fatta per tutte le variabili oggetto di indagine, soltanto dopo aver effettuato una sperimentazione sul campo dello schema stesso.

La descrizione riportata riguarda gli aspetti essenziali dello schema di campionamento areale utile per un'indagine sulla popolazione, ovvero che fa riferiemnto alle famiglie come unità di rilevazione. L'utilità della trattazione risiede nel fatto che attraverso i pochi elementi forniti si può cercare di avere un'idea della validità dei risultati campionari, se disponiamo di alcune informazioni sulla variabilità dei caratteri oggetto di indagine.

8. Stime ed errori campionari

Dopo aver delineato uno schema di campionamento areale da applicare in ciascuna provincia, facendo riferimento ad una provincia con cinque comuni autorappresentativi (con più di 20.000 abitanti) e due strati di sezioni distinti in base alla dimensione dei comuni di appartenenza (con meno di 20.000 abitanti), prendiamo in esame il problema del calcolo delle stime e degli errori campionari, che si hanno attraverso l'applicazione dello schema stesso.

All'interno di ogni strato s (s=1, 2,..., S) di sezioni e di ogni comune autorappresentativo la stima

$$n_S = \frac{A_S}{a_S} \sum_{j=1}^{a_S} \frac{A_{sj}}{a_{sj}} \sum_{k=1}^{a_{sj}} N_{sjk} \qquad (37)$$

del numero N_S di individui con una determinata caratteristica è funzione dei dati campionari raccolti nelle a_S sezioni campione e negli a_{sj} segmenti. Mentre la varianza (omettendo i fattori di correzione dovuti all'estrazione senza ripetizione)

$$\mathrm{var}(n_S) = \frac{A_S}{f_1} \sigma^2_{sj} + \frac{A_S}{a_S} \sum_{j=1}^{A_S} \frac{A^2_{sj}}{a_{sj}} \sigma^2_{sjk} \qquad (38)$$

della stima n_s dipende dalla varianza σ^2_{sj} del numero N_{sj} di individui tra sezioni di censimento e dalla varianza σ^2_{sjk} del numero N_{sjk} di individui tra segmenti. Tuttavia, l'espressione indicata per la varianza delle stime n_s può essere scritta diversamente

$$\text{var}(n_s) = A_s \frac{1}{f_1} \overline{A}_s . F_o \text{ var } (N_{sjkl}) + A_s \overline{A}_s . \frac{1}{f_1} \frac{1}{f_2} F_o \text{ var } (N_{sjkl}) \qquad (39)$$

in funzione della varianza del numero N_{sjkl} di individui per famiglia, del numero medio $\overline{A}_s$ dei segmenti per sezione e della dimensione F_o dei segmenti, oppure nella forma seguente,

$$\text{var}(n_s) = \sigma^2_c D^2 = \sigma^2_c [\, 1 + \rho \, (F_o - 1)] \qquad (40)$$

in base alla varianza σ^2_c relativa ad un campione casuale semplice ed all'effetto D^2 prodotto da un particolare schema di campionamento, che prevede l'impiego dei segmenti delle unità di rilevazione con una dimensione F_o. Pertanto, al fine di calcolare la varianza delle stime n_s, nel caso dell'espressione (39) sarebbe necessario conoscere la varianza del numero N_{sjkl} di individui per famiglia nello strato considerato. Invece, volendo ricorrere all'espressione (40) è necessario conoscere la varianza σ^2_c e l'effetto D^2. Come accade usualmente, prima di effettuare la rilevazione e di fare una stima dell'errore campionario attraverso i risultati della rilevazione, una valutazione approssimata di tale errore può essere fatta in base alle informazioni disponibili sulla popolazione oggetto di indagine e sulle esperienze fatte precedentemente per lo stesso tipo di rilevazioni.

Per quanto riguarda i dati disponibili sulle famiglie, tra le variabili interessanti possiamo prendere in esame quella relativa alle distribuzioni delle famiglie secondo il numero di occupati al censimento del 1981. Dalla lettura dei dati contenuti nella tabella 2 si vede che la varianza del numero N_l di occupati per famiglia di quasi tutte le province è inferiore a uno e per semplicità faremo riferimento a tale valore.

Inoltre, per fare una valutazione a priori, prima cioè dell'uso dello schema ipotizzato di campionamento areale, consideriamo mediamente per provincia cinque strati di sezioni con 20.000 famiglie corrispondenti ai cinque comuni autorappresentativi, uno strato di sezioni con 53.000 famiglie appartenenti ai comuni di media dimensione ed uno strato di sezioni con 37.000 famiglie appartenenti ai piccoli comuni. Allora, con una frazione di campionamento del 7‰, si ha un campione di 140 famiglie per ogni comune

Tab.2 - Distribuzioni delle famiglie secondo il numero di occupati; valori medi, varianza ed errore quadratico medio del numero di occupati per famiglia. Dati del censimento 1981 per province e regioni.

Province **Regioni**	numero occupati 0	1	2	3	4	≥5	Totale	Media	Var.	E.Q.M.
011 Torino	253072	336181	234245	29908	5595	974	859975	1.072	.777	.882
012 Vercelli	55286	45338	43015	8006	1607	255	153507	1.062	.966	.983
013 Novara	63174	63315	50657	8370	1787	345	187648	1.058	.897	.947
014 Cuneo	59969	72811	50926	10333	2725	688	197452	1.114	.938	.968
015 Asti	27934	27487	19904	3831	895	182	80233	1.038	.922	.960
016 Alessandria	68295	66129	40347	6480	1254	226	182731	.944	.815	.903
010 **Piemonte**	527730	611261	439094	66928	13863	2670	1661546	1.059	.840	.917
020 **V.Aosta**	14078	17879	9898	1400	291	55	43601	.993	.759	.871
031 Varese	72291	97791	81373	14397	3454	616	269922	1.188	.899	.948
032 Como	67544	101502	73601	13944	3420	743	260754	1.181	.886	.941
033 Sondrio	15598	23385	13591	2621	741	217	56153	1.113	.874	.935
034 Milano	370630	567155	417518	57650	11199	1968	1426120	1.143	.787	.887
035 Bergamo	69083	118190	82078	17399	4558	1135	292443	1.226	.906	.952
036 Brescia	86910	140345	89937	17953	4196	955	340296	1.163	.860	.927
037 Pavia	66978	67570	50408	7992	1419	232	194599	1.024	.852	.923
038 Cremona	34355	44003	32204	5778	1252	262	117854	1.121	.878	.937
039 Mantova	32039	45444	35985	7903	1842	433	123646	1.218	.948	.974
030 **Lombard.**	815428	1205385	876695	145637	32081	6561	3081787	1.154	.842	.918
041 Trento	45582	63809	34232	5872	1602	361	151458	1.044	.813	.902
042 Bolzano	29519	56099	30789	8622	3589	1962	130580	1.284	1.133	1.064
040 **Tr.A.A.**	75101	119908	65021	14494	5191	2323	282038	1.155	.976	.988
051 Verona	65826	105215	65708	13084	3150	687	253670	1.151	.857	.926
052 Vicenza	53981	86585	69447	14357	3853	890	229113	1.259	.936	.967
053 Belluno	26088	29458	18518	3089	617	115	77885	1.012	.826	.909
054 Treviso	52861	89384	63293	13863	3641	892	223934	1.235	.921	.960
055 Venezia	66702	123141	61416	10776	2201	476	264712	1.094	.742	.861
056 Padova	52570	102747	67442	15007	3849	960	242575	1.248	.887	.942
057 Rovigo	23634	31433	22473	4311	1094	228	83173	1.140	.904	.951
050 **Veneto**	341662	567963	368297	74487	18405	4248	1375062	1.180	.870	.933
061 Pordenone	25885	34746	25507	4463	938	189	91728	1.132	.862	.929
062 Udine	55680	72151	46617	8100	1672	300	184520	1.072	.832	.912
063 Gorizia	19162	21533	12104	1602	295	35	54731	.948	.744	.863
064 Trieste	49914	42527	23176	2854	388	36	118895	.834	.717	.847
060 **Fr.V.G.**	150641	170957	107404	17019	3293	560	449874	1.006	.811	.900

[segue tab.2]

Province **Regioni**	numero occupati 0	1	2	3	4	≥5	Totale	Media	Var.	E.Q.M.
071 Imperia	35662	31442	18321	2704	486	87	88702	.886	.783	.885
072 Savona	46059	45215	23221	3147	527	77	118246	.876	.728	.853
073 Genova	167885	172377	74639	7624	1088	174	423787	.825	.643	.802
074 La Spezia	39069	38671	16370	1542	167	23	95842	.802	.623	.789
070 **Liguria**	288675	287705	132551	15017	2268	361	726577	.838	.672	.820
081 Piacenza	32382	40512	24759	4161	806	155	102775	1.036	.811	.901
082 Parma	45876	52866	39254	6844	1335	235	146410	1.082	.869	.932
083 Reggio E.	38986	47769	45325	9183	1995	472	143730	1.227	.968	.984
084 Modena	48251	62212	71444	15571	3585	746	201809	1.337	1.007	1.003
085 Bologna	101771	112182	106936	19589	3936	763	345177	1.172	.940	.970
086 Ferrara	40650	46183	37648	6909	1346	267	133003	1.120	.905	.952
087 Ravenna	34950	43176	37755	7678	1785	437	125781	1.201	.965	.982
088 Forlì	53782	78446	54929	10312	2421	588	200478	1.157	.879	.938
080 **E.Rom.**	396648	483346	418050	80247	17209	3663	1399163	1.177	.966	.933
091 Massa Carr.	28546	32141	13332	1451	187	31	75688	.846	.635	.797
092 Lucca	44492	51667	32028	5195	982	165	134529	1.011	.808	.899
093 Pistoia	25195	31168	26347	5487	1226	210	89633	1.186	.946	.973
094 Firenze	106209	149714	121608	23542	5093	1072	407238	1.201	.905	.951
095 Livorno	41688	53311	25368	3102	429	70	123968	.931	.676	.822
096 Pisa	38603	49081	36760	6358	1239	276	132317	1.119	.865	.930
097 Arezzo	28629	35135	31991	5901	1394	357	103407	1.201	.951	.975
098 Siena	25854	31506	24106	4507	897	232	87102	1.125	.900	.949
099 Grosseto	28358	34466	15760	2092	309	61	81046	.911	.686	.828
090 **Toscana**	367574	468189	327300	57635	11756	2474	1234928	1.097	.858	.926
101 Perugia	53226	69096	53500	9552	2074	512	187960	1.147	.892	.945
102 Terni	25052	33633	16035	1992	324	73	77109	.951	.687	.829
100 **Umbria**	78278	102729	69535	11544	2398	585	265069	1.090	.840	.917
111 Pesaro-Urb.	31875	39751	30289	5340	1267	265	108787	1.128	.895	.946
112 Ancona	40863	51996	41397	7358	1783	427	143824	1.155	.912	.955
113 Macerata	23471	30594	28039	6736	1920	560	91320	1.285	1.065	1.032
114 Ascoli Pic.	25554	39808	31305	7880	2156	487	107190	1.279	1.000	1.000
110 **Marche**	121763	162149	131030	27314	7126	1739	451121	1.204	.965	.982

[segue tab.3]

Province Regioni	numero componenti 1	2	3	4	5	6	≥7	Totale	Media	Var	E.Q.M
071 Imperia	23926	25010	19792	14399	4061	1066	448	88702	2.489	1.634	1.278
072 Savona	30706	33753	27959	18890	5276	1202	460	118246	2.490	1.556	1.248
073 Genova	119369	117748	97587	67531	16520	3566	1466	423787	2.435	1.532	1.238
074 La Spezia	23905	27198	23198	16689	3870	743	239	95842	2.505	1.482	1.217
070 **Liguria**	197906	203709	168536	117509	29727	6577	2613	726577	2.460	1.543	1.242
081 Piacenza	21998	28068	25752	18657	5893	1626	781	102775	2.673	1.697	1.303
082 Parma	29561	40984	36849	26533	8638	2650	1195	146410	2.702	1.710	1.308
083 Reggio E.	25503	37545	37214	27835	10379	3519	1735	143730	2.844	1.843	1.357
084 Modena	29953	53022	54670	40552	15671	5449	2492	201809	2.927	1.803	1.343
085 Bologna	69397	97944	93081	58590	18511	5264	2390	345177	2.665	1.608	1.268
086 Ferrara	21548	36617	35357	25977	9200	3071	1233	133003	2.841	1.726	1.314
087 Ravenna	23043	32672	32447	24924	8507	2737	1451	125781	2.819	1.809	1.345
088 Forlì	33049	47659	49532	45063	16836	5731	2608	200478	2.963	1.922	1.386
080 **E.Rom.**	254052	374511	364902	268131	93635	30047	13885	1399163	2.799	1.765	1.328
091 Massa Carr.	16367	20309	18021	15313	4366	951	361	75688	2.674	1.649	1.284
092 Lucca	25208	34494	32415	26951	10615	3477	1369	134529	2.845	1.893	1.376
093 Pistoia	14294	23330	22349	18321	7612	2677	1050	89633	2.931	1.894	1.376
094 Firenze	69744	101878	101084	83721	34377	11971	4463	407238	2.914	1.911	1.382
095 Livorno	22152	34967	30629	26233	7524	1854	609	123968	2.758	1.606	1.267
096 Pisa	21493	34686	32889	27854	10427	3624	1344	132317	2.904	1.837	1.356
097 Arezzo	16089	25423	26244	21124	9189	3624	1714	103407	2.996	2.009	1.417
098 Siena	14658	23508	22151	15854	7085	2651	1195	87102	2.884	1.931	1.389
099 Grosseto	15594	23646	20419	14834	4548	1352	653	81046	2.701	1.659	1.288
090 **Toscana**	215599	322241	306201	250205	95743	32181	12758	1234928	2.867	1.854	1.362
101 Perugia	30106	44273	43323	41194	18215	7436	3413	187960	3.048	2.127	1.459
102 Terni	11722	20808	18833	17101	5885	1977	783	77109	2.918	1.792	1.338
100 **Umbria**	41828	65081	62156	58295	24100	9413	4196	265069	3.010	2.033	1.426
111 Pesaro-Urb.	16741	26066	25490	24587	10476	3770	1657	108787	3.036	2.021	1.422
112 Ancona	22593	35608	34065	32948	12509	4347	1754	143824	2.981	1.920	1.386
113 Macerata	12555	21106	20966	20603	10000	4168	1922	91320	3.160	2.167	1.472
114 Ascoli Pic.	14113	22619	23641	25938	12979	5407	2493	107190	3.254	2.226	1.492
110 **Marche**	66002	105399	104162	104076	45964	17692	7826	451121	3.095	2.079	1.442

[segue tab.3]

Province Regioni	numero componenti 1	2	3	4	5	6	≥7	Totale	Media	Var	E.Q.M.
181 Cosenza	35553	47162	42244	48868	28696	12167	9289	223979	3.275	2.653	1.629
182 Catanzaro	39609	46079	37766	44453	29094	13890	11508	222399	3.292	2.947	1.717
183 Reggio C.	30661	39311	31182	35957	22162	9558	7104	175935	3.209	2.720	1.649
180 **Calabria**	105823	132552	111192	129278	79952	35615	27901	622313	3.263	2.778	1.667
191 Trapani	21435	31299	26245	31583	15395	5081	2890	133928	3.112	2.244	1.498
192 Palermo	55971	77332	69771	84073	47047	17949	12035	364178	3.250	2.485	1.576
193 Messina	44147	52257	42298	47914	23134	7902	4617	222269	2.981	2.313	1.521
194 Agrigento	24131	32114	26407	31329	18409	7591	4504	144485	3.198	2.549	1.597
195 Caltaniss.	14709	18758	16179	18866	11479	4813	2906	87710	3.225	2.597	1.611
196 Enna	11709	14453	11294	12760	6941	2731	1702	61590	3.061	2.491	1.578
197 Catania	49613	66245	61264	72831	38258	14022	8802	311035	3.197	2.397	1.548
198 Ragusa	15636	21705	18788	21259	9392	2689	1276	90745	3.003	2.068	1.438
199 Siracusa	20132	27239	24284	30335	14846	4782	2697	124315	3.142	2.262	1.504
190 **Sicilia**	257483	341402	296530	350950	184901	67560	41429	1540255	3.151	2.401	1.549
201 Sassari	20993	25332	24644	27910	16118	7442	6080	128519	3.307	2.753	1.659
202 Nuoro	14455	14601	13444	14858	10180	5400	5815	78753	3.396	3.267	1.808
203 Oristano	7981	9639	8625	9362	5651	2731	2421	46410	3.279	2.871	1.694
204 Cagliari	29962	38021	39746	47394	27873	12920	11818	207734	3.439	2.802	1.674
200 **Sardegna**	73391	87593	86459	99524	59822	28493	26134	461416	3.379	2.879	1.697
000 **ITALIA**	3323456	4402980	4117217	4008008	1773621	628719	378336	18632337	2.995	2.174	1.475

$$\text{var}\,(n_r) = 97.709.400 \times 5 = 488.545.000 < 22.200^2 \qquad (54)$$

della stima risulta inferiore a 500 milioni e corrisponde ad un errore quadratico medio di circa 22.200 unità.

Infine, se passiamo a considerare la precisione dei risultati che ci possiamo attendere in base ad un campione casuale semplice della stessa dimensione del campione areale ed uno schema particolare di campionamento con un effetto $D^2 = 2$, si vede facilmente che, per un campione casuale semplice con ripetizione, la varianza

$$\text{var}\,(n_r') \cong \frac{1.000.000}{0,007} \times 3 \cong 429.000.000 \qquad (55)$$

della stima n_r' del numero di componenti delle famiglie della regione r è inferiore a 429 milioni. Mentre la varianza

$$\text{var}(n_r) = \text{var}(n_r')\, D^2 = 429.000.000 \times 2 = 858.000.000 < 29.500^2 \qquad (56)$$

della stima n_r, che è ottenuta in funzione dello schema particolare di campionamento, è pari a 858 milioni e corrisponde ad un errore quadratico medio di circa 29.500 unità.

Da questi risultati ci sembra di poter dire che un campione areale possa fornire dei risultati affidabili e abbastanza precisi in relazione a diverse variabili, anche se a posteriori, quando cioè sono disponibili i dati raccolti sul campo mediante il campione, è possibile (Kalton, 1979; Kish e Frankel, 1970) calcolare il valore effettivo della varianza delle stime, includendo l'effetto delle diverse fonti di errore connesse con la rilevazione statistica.

Un altro aspetto interessante per l'argomento trattato riguarda il confronto tra i risultati forniti in questo contesto e quelli che si hanno attraverso l'uso del campione Istat sulle forze di lavoro. Da questo punto di vista notiamo che per il campione usato fino all'anno trascorso (Istat, 1978) è stato determinato lo scarto teorico delle stime nazionali e regionali. In particolare, per una regione con un milione di occupati, dei quali 750.000 di sesso maschile e 250.000 di sesso femminile, la stima dell'errore quadratico medio teorico,

$$S_{M+F} = \left(S_M^2 + S_F^2 + 2\delta_{M,F}\, S_M\, S_F\right)^{1/2} \cong 14.000 \qquad (57)$$

in base alle varianze S_M^2, S_F^2 e alla covarianza $\delta_{M,F}\, S_M\, S_F$ (Istat, 1978, pp.44, 45, 61) risulta di circa 14.000 unità, che è un valore intermedio a quelli relativi alle due stime a priori (13.000 e 17.000) dell'errore quadratico medio per il campione ipotetico di tipo areale . Pertanto, si può pensare che lo schema di campionamento areale rappresenti una valida soluzione alternativa per l'attività di documentazione statistica connessa a diverse rilevazioni ufficiali periodiche.

Naturalmente, le considerazioni fatte devono essere sottoposte ad una verifica empirica sul campo, ma riteniamo che valga la pena di approfondire l'argomento, affrontando tutti i problemi di natura pratica e teorica che si devono risolvere per mettere a punto uno schema adeguato di campionamento areale.

9. Considerazioni conclusive

Gli elementi da definire per la formazione di un campione areale da utilizzare per alcune rilevazioni ufficiali in Italia sono molteplici e la trattazione riportata nei paragrafi precedenti riguarda principalmente alcuni degli aspetti principali relativi allo schema di campionamento. Tuttavia riteniamo che quanto è emerso sia sufficiente a far riflettere sull'opportunità di procedere nello studio e nella sperimentazione sui varî elementi che possono contribuire al raggiungimento dei risultati utili per lo svolgimento di una efficiente attività di documentazione statistica.

Gli aspetti pratici dell'argomento trattato che devono essere affrontati in maniera adeguata sono diversi e riguardano, ad esempio, il fatto che verosimilmente tutti i comuni non si sono attenuti in maniera uniforme alle indicazioni fornite dall'Istat sulla formazione delle sezioni di censimento. Inoltre, il problema della rilevazione delle nuove unità è stato oggetto soltanto di qualche cenno, mentre, come è facile intuire, esso riveste una notevole importanza quando si desidera che il campione sia in grado di mettere in luce tempestivamente i mutamenti che interessano il campo dell'indagine statistica. Infine, la definizione dei segmenti deve essere stabilita in maniera sufficientemente precisa per fare in modo che ogni unità non possa essere attribuita a più di un segmento. In proposito, desideriamo rilevare che tutti questi elementi hanno un certo effetto sui risultati della rilevazione, ma solo una sperimentazione sul campo può dare informazioni affidabili sulle scelte più opportune.

Oltre che sul piano pratico, anche dal punto di vista metodologico è necessario affrontare in maniera deguata alcuni argomenti che hanno un ruolo fondamentale rispetto all'impiego dei risultati che si hanno attraverso l'applicazione di uno schema di campionamento areale. In questo ambito, ad esempio, rientra la messa a punto dei procedimenti per la misura degli errori di rilevazione di natura diversa dagli errori campionari; il rinnovo parziale delle unità di osservazione per le indagini statistiche periodiche e l'analisi dei procedimenti di stima per individuare gli stimatori più efficienti.

Volendo fare qualche considerazione sommaria sull'opportunità di usare un determinato schema di campionamento, oltre la comparabilità degli errori campionari connessi al campione tradizionale usato dall'Istat per la rilevazione sulle forze di lavoro con quelli dello schema ipotizzato di campionamento areale, si può dire che il primo tipo di campione comporta la raccolta di dati presso unità di rilevazione che sono territorialmente distanti l'una dall'altra. Invece, il campione areale comporta la raccolta di dati presso unità statistiche che a gruppi risultano generalmente vicine e facilitano le operazioni sul campo. Tale circostanza può avere diverse conseguenze e, in particolare, incide sulla variabilità degli stimatori e sui costi di raccolta dei dati.

Più in generale, ci sembra di dover dire che non è possibile stabilire in maniera astratta la convenienza di usare un determinato schema di campione areale rispetto ad altri schemi, eventualmente predisposti attraverso il ricorso a liste più o meno buone delle unità di rilevazione o di osservazione, in quanto ogni tipo di campione dovrebbe essere messo a punto in rapporto alla situazione concreta particolare in cui viene impiegato, la quale risulta caratterizzata anche dal sistema di rilevazioni campionarie e complete, che concorrono allo svolgimento dell'attività di documentazione statistica. In altre parole, la definizione dei metodi di indagine sul piano tecnico ed operativo dipende da un insieme abbastanza ampio di elementi, che in parte possono essere acquisiti soltanto attraverso una vasta sperimentazione estesa nel tempo e che, al limite, coinvolgono anche l'attività futura non ancora definita esattamente dall'ente responsabile.

A questo livello la programmazione delle indagini per la raccolta di dati statistici diventa entro certi limiti una scelta di natura politica e può essere orientata in un senso particolare, ma non può essere definita esclusivamente attraverso degli elementi di natura strettamente tecnica. Pertanto, senza voler trarre delle conclusioni definitive sull'opportunità di adottare un tipo di campione piuttosto di un altro, si possono fare delle valutazioni su alcune caratteristiche dei diversi schemi di campionamento.

Per quanto riguarda una parte delle operazioni necessarie per selezionare le unità di rilevazione, notiamo che talvolta si incontrano gli stessi problemi per diversi tipi di campioni, come accade, per esempio, quando si ricorre al criterio della stratificazione. Tuttavia, nella fase dell'estrazione delle unità di rilevazione all'interno delle unità appartenenti al penultimo stadio di campionamento, cioè quello delle sezioni, in teoria il campione areale presuppone la possibilità di costruire un elenco delle unità di rilevazione attraverso un microcensimento, al fine di aggiornare i dati di sezione e di suddividere le unità interessate in segmenti, alcuni dei quali saranno selezionati ed inseriti esaustivamente nel campione.

Naturalmente, l'attività in questione comporta un certo impegno, ma permette anche di ricavare delle informazioni utili per controllare, aggiornare ed uniformare nel tempo le caratteristiche delle sezioni di censimento nelle quali il territorio è stato suddiviso. Inoltre, notiamo che quando una sezione è troppo ampia ed il microcensimento rappresenta un aggravio di costi, si può pensare di ricorrere ad una suddivisione approssimativa della sezione in tante sottosezioni della dimensione media desiderata, in modo da avvicinarsi alla situazione ideale per l'impiego di un campione areale.

In relazione alla definizione appropriata ed all'uso delle sezioni di censimento, ricordiamo che anche in altre occasioni è stata sottolineata la necessità di un riferimento spaziale per i dati statistici più adeguato alle esigenze conoscitive, quale la sezione di censimento, da utilizzare anche nell'ambito delle rilevazioni campionarie (Marbach, 1990), oppure per raccogliere informazioni sulle forze di lavoro presenti o sulle famiglie di fatto (Fabbris, 1990). Pertanto, oltre l'efficienza delle stime in rapporto ai costi ed alla precisione, vale la pena mettere in evidenza il fatto che un campione areale è in grado di fornire delle informazioni aggiuntive rispetto a quelle che si ricavano attraverso altri tipi di campione.

Riferimenti bibliografici

DREW J.D., SINGH G.H., CHOUDHRY G.H. (1982) - Evaluation of Small Area Estimation Techniques for Canadian Labour Force Survey. *Survey Methodology*, vol.8, pp.17-47

ESPANA E.G. (1975) - *Design of the General Population Survey*. Instituto Nacional de Estadistica, Madrid

FABBRIS L. (1990) - Problemi metodologici nell'impiego dell'indagine campionaria per la produzione di statistiche ufficiali. In : *Atti del Convegno "La Statistica Italiana per l'Europa del 1993"*, Roma, 21-23 maggio 1990

FELLEGI I.P. (1964) - Reponse Variance and its Estimation. *JASA*, n.59, pp.1016-1041

FELLEGI I.P., GRAY G.B., PLATEK R. (1967) - The new Design of the Canadian Labour Force Survey. *JASA*, n.62, pp.421-453

FSO - FEDERAL STATISTICAL OFFICE (1969) - The German Microcensus. Statistiches Bundesamt, *Studies on Stat.*, n.23

GINI C.(1927) - Une application de la méthode représentative aux materiaux du dernier recensement de la population italienne (1er décembre 1921). *Bulletin de l'Institut International de Statistique*, tome XXIII

GIUSTI F.(a cura di) (1973) - *Considerazioni sul campionamento per aree*. Documento di lavoro non pubblicato, n.41, ISTAT

GOODMAN R., KISH L. (1950) - Controlled Selection, a Technique in Probability Sampling. *JASA*, 45, pp.350-372

HANSEN M.H., HURWITZ W.N., BERSHAD M.A. (1961) Measurement Errors in Censuses and Survey, *Bull. of ISI*, 38, pp.359-374

HANSEN M.H., HURWITZ W.N., MADOW W.G.(1953). *Sample Survey Methods and Theory*. Wiley, New York

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

ISTAT (1991) - *Compendio Statistico Italiano*, ed.1991, Roma

KALTON G. (1979) - Ultimate Cluster Sampling. *JRSS*, A, 142, pp.210-222

KHAMIS S.H., ALONZO D.P. (1975) - Changes in Methods, Scope and Concepts in the 1980 World Census of Agriculture. *Bull. of the Int.Stat. Institute*, 46, pp.54-82

KISH L. (1965) - *Survey Sampling*, Wiley, New York

KISH L. (1979) - Samples and Censuses, *Int.Stat.Rev.*, n.47, pp.99-109

KISH L. (1987) - *Statistical Design for Research*, Wiley, New York

KISH L., FRANKEL M.R. (1970) - Balanced Repeated Replications for Standard Errors. *JASA*, 65, pp.1071-1094

KISH L., FRANKEL M.R. (1974) - Inference from Complex Samples. *JRSS*, B, 36, pp.1-37

KISH L., VERMA V. (1983) - Censuses and Samples: Combined Uses and Designs. *Bull. of the Int.Stat.Institute*, 44, pp.66-82

MADOW W., NISSELSON H., OLKIN I. (eds) (1983) - *Incomplete Data in Sample Surveys*. Academic Press, New York

MARBACH G. (1990) - Ricerche campionarie e microaree. In : *Atti del Convegno "La Statistica Italiana per l'Europa del 1993"*, Roma, 21-23 maggio 1990

O'MUIRCHEARTAIGH C. A. (1982) - Methodology of the Response Error Project. *WFS Sc.Rep.* n.28, London

O'MUIRCHEARTAIGH C. A., WONG S.T. (1981) - The Impact of Sampling Theory on Survey Practice: a Review. *Bull. of the ISI*, 49, 1, pp. 465-493

OPCS - OFFICE OF POPULATION CENSUS AND SURVEYS (1973). *The General Household Survey*. London

PURCELL N.J., KISH L. (1980) - Postcensal Estimates for Local Areas (or Domains). *ISI*, 48, pp.3-18

REDFERN P.(1974) - The Different Roles of Population Censuses and Interview Surveys, particularly in the U.K. Context. *Int.Stat.Review*, vol.42, n.2, pp.131-146

REDFERN P.(1989) - Population Registers: Some Administrative and Statistical Pros and Cons. *JRSS*, A, 152, pp.1-41

SARNDAL C.E. (1984) - Design-Consistent Versus Model-Dependent Estimation for Small Domains. *JASA*, 79, pp.624-631

STATISTICS CANADA (1976) - *Methodology of the Canadian Labour Force Survey.* Min.of Ind.Trade and Commerce, Ottawa

U.S. CENSUS BUREAU (1978) - *The Current Population Survey: Design and Methodology*. Techn.Paper n.40, Washington, DC, Gov.Printing Office

VERMA V., SCOTT C., O'MUIRCHEARTAIGH (1980) - Sample Designs and Sampling Errors for the World Fertility Survey. *JRSS*, A, 143, pp. 431-473