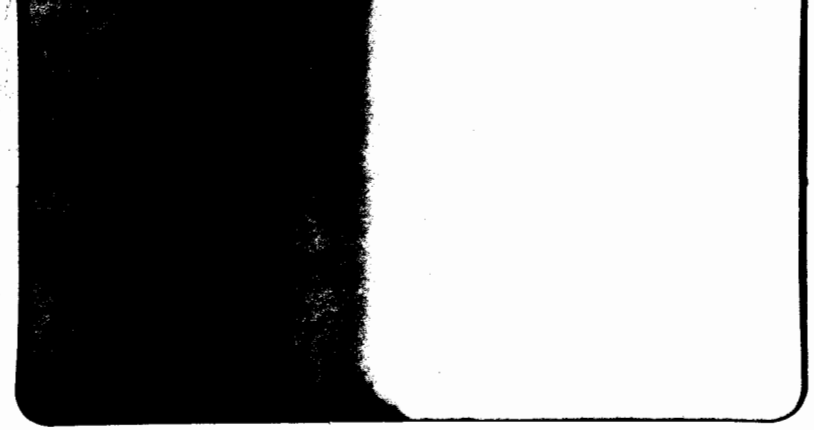

I lavori raccolti in questa collana hanno avuto origine nell'ambito del progetto di ricerca dell'ISTAT «Le statistiche dei consumi privati nel sistema statistico nazionale» e del progetto di ricerca MURST 40% «La misura dei consumi privati: uno studio sull'accuratezza, coerenza e qualità dei dati».
Al progetto di ricerca hanno partecipato i ricercatori dell'ISTAT e dei seguenti Dipartimenti e Istituti universitari:

- Dipartimento di Scienze Statistiche, Bologna
- Dipartimento Statistico, Firenze
- Dipartimento di Scienze Statistiche, Perugia
- Dipartimento di Contabilità Nazionale, Roma
- Istituto di Statistica e Matematica, Istituto Universitario Navale, Napoli
- Istituto di Statistica, Messina.

UN METODO DI STIMA GENERALIZZATO PER LE INDAGINI SULLE FAMIGLIE E SULLE IMPRESE

Piero Demetrio Falorsi* Stefano Falorsi*

Rapporto di ricerca n. 13

CON PRI - La misura dei consumi privati

*** Servizio Studi Metodologici - Istituto Nazionale di Statistica**

**Dipartimento di Scienze Statistiche "Paolo Fortunati"
dell'Università degli Studi di Bologna
Gennaio 1995**

INDICE

Premessa

Il problema principale nella scelta dello stimatore per le indagini campionarie concrete su larga scala è quello di individuare un metodo di stima che risponda a:

- **criteri di efficenza delle stime:** i) *in termini di bassa varianza delle stime*; *ii) riduzione della distorsione delle stime:* in presenza di fenomeni distorsivi quali il fenomeno delle "mancate risposte totali e parziali", ed il fenomeno della sottocopertura della lista di estrazione del campione, rispetto alle popolazioni oggetto di indagine;
- **criteri di coerenza esterna ed interna delle stime:** i) *il problema di coerenza esterna delle stime*: nasce ogniqualvolta si dispone, da fonti esterne, di totali noti aggiornati sulla popolazione oggetto di indagine. Le stime prodotte dall'indagine dei totali noti devono , in generale, coincidere o non discostarsi molto, dal valore noto di tali totali.; ii) *il problema di coerenza interna delle stime*: nasce dall'esigenza di produrre stime di uno stesso aggregato che siano coerenti tra loro. Tutte le stime prodotte dall'indagine di uno stesso aggregato devono coincidere tra loro. Questo risultato può essere ottenuto utilizzando un unico sistema di pesi per il riporto dei dati all'universo; invece, l'uso, di sistemi di riporto all'universo non unici, porta a problemi di coerenza interna delle stime.

Il metodo di stima qui presentato rientra nell'ambito degli *stimatori di ponderazione vincolata* e risponde ai criteri appena

elencati. Esso viene applicato, attualmente, per il calcolo dei coefficenti di riporto all'universo di alcune importanti indagini campionarie dell'Istituto Nazionale di Statistica sulla popolazione, quali ad esempio: l'"Indagine multiscopo sulle famiglie" L'"indagine sulle vacanze degli italiani", l'"Indagine sugli sbocchi professionali dei laureati" ecc..; inoltre si sta studiando la possibilità di applicare tale metodo di stima ad altre fondamentali indagini campionarie dell'Istituto quale ad esempio l'Indagine sui consumi di famiglia", l'"Indagine sulle forze di lavoro" . Ricordiamo, inoltre, che a partire dal 1994, questo metodo, è stato utilizzato per produrre stime di alcune importanti indagini dell'Istat sulle imprese, quali ad esempio: l'indagine sulla struttura delle aziende agricole ; l"indagine sul mercato del lavoro" condotta dal Ministero del Lavoro ecc.

Nelle pagine successive, oltre alla descrizione formale dello stimatore e delle sue principali proprietà statistiche (cfr. cap 2.), viene presentata una procedura informatica di tipo *user friendly* che permette di applicare facilmente il metodo in oggetto alla maggiorparte delle indagini campionarie dell'Istat (cfr. cap 3.). In particolare nel paragrafo 3.7 sono descritte le modalità di applicazione della procedura informatica alle indagini sulla popolazione e sulle imprese.

1. Introduzione

Ogni indagine campionaria condotta su larga scala ha, generalmente, la finalità di fornire un elevato numero di stime di parametri della popolazione, che possono essere di tipo differente (frequenze assolute, totali, proporzioni, medie, ecc.).

Poiché, indipendentemente dal metodo di stima adottato, le stime di frequenze assolute, di proporzioni o di medie si possono ricavare da quella di un totale, limiteremo la descrizione soltanto a quest'ultimo tipo di stima.

Il principio su cui è fondato qualsiasi metodo di stima campionaria è quello che il sotto insieme delle unità della popolazione incluse nel campione deve rappresentare anche il sotto insieme complementare costituito dalle rimanenti unità della popolazione stessa (ISTAT, 1989; Statistics Canada 1976). Tale principio viene realizzato attribuendo a ciascuna unità inclusa nel campione un peso, che può essere visto come numero di numero di elementi della popolazione rappresentati da detta unità.

Se, ad esempio, ad una unità campionaria viene attribuito un peso pari a 50, ciò indica che tale unità rappresenta se stessa ed altri 49 elementi della popolazione che non sono stati sottoposti ad indagine.

In generale, per ottenere la stima di un totale (ad esempio il reddito totale) si devono eseguire le tre seguenti operazioni:

(i) determinare il peso da attribuire a ciascuna unità inclusa nel campione;

(ii) moltiplicare il valore relativo ad un data variabile oggetto di indagine, rilevata sulla generica unità inclusa nel campione, per il peso attribuito alla medesima unità (nell' esempio in questione, il reddito di ciascun individuo campionato viene moltiplicato per il corrispondente peso);

(iii) effettuare la somma dei prodotti di cui al punto (ii).

Nelle indagini effettive, generalmente basati su disegni di campionamento complessi, il peso da attribuire a ciascuna unità è ottenuto in base ad una procedura articolata in più passi (Bureau of the Census, 1978):

(a) in primo luogo, viene calcolato un peso iniziale, definito *peso diretto*, determinato in funzione del disegno di campionamento come reciproco della probabilità di inclusione dell' unità campionata;

(b) successivamente, vengono calcolati dei fattori correttivi del peso base, che possono essere distinti in: fattori
-per mancata risposta totale;
-che consentono di rispettare la condizione di uguaglianza tra alcuni parametri noti della popolazione e le corrispondenti stime campionarie;

(c) infine, viene determinato un peso, noto sotto il nome di *peso finale*, espresso come prodotto del peso base per i fattori correttivi.

Il presente lavoro è finalizzato ad illustrare le principali caratteristiche di un metodo per l'ottenimento dei pesi finali che può essere adottato per le indagini sulle imprese e sulle famiglie in tutti i casi in cui si disponga di totali noti sulla popolazione oggetto

d'indagine o sull' archivio da cui il campione è stato estratto. Il metodo in oggetto porta alla determinazione dei pesi finali che, sotto determinate ipotesi (Binder, 1988) sono correttivi della mancata risposta e consentono di rispettare il vincolo dell' uguaglianza tra i totali noti e le corrispondenti stime campionarie. Le proprietà statistiche del metodo, che determina i pesi finali in modo che essi risultino il più vicino possibile (sulla base di una metrica prescelta) ai pesi base, verranno descritte nel capitolo 2. Il capitolo 3 illustra le una procedura informatica scritta in linguaggio SAS che permette di calcolare i pesi mediante il metodo in oggetto. Gli elementi fondamentali della procedura sono:

- generalità di applicazione: può essere applicata alla maggior parte delle indagini campionarie ISTAT sulle famiglie e sulle imprese;
- possibilità di definire differenti domini di studio;
- facilità di utilizzazione: può essere utilizzata direttamente dall'utente che è responsabile dell'indagine, poichè l'applicazione della procedura per le singole indagini campionarie dell'Istituto richiede solo poche e semplici modifiche ai programmi di cui essa è composta;
- parametrizzazione: la procedura dipende da un'insieme di parametri che l'utente deve specificare. Al variare di tali parametri si modifica il vettore dei pesi finali che si ottiene;
- ricchezza di informazioni statistiche sul vettore dei pesi finali ottenuto.

2. Procedura generale di stima

2.1. Premessa

Indichiamo con: U, l'insieme delle unità appartenenti alla popolazione oggetto di indagine; U_L, l'insieme delle unità presenti sulla lista (o sulle liste) da cui si seleziona il campione s* ($s^* \subseteq U_L$) mediante il disegno p(•); n*, il numero di unità appartenenti all'insieme s*; s ($s \subseteq s^*$), l'insieme delle unità campionarie rispondenti; n, il numero di unità appartenenti all' insieme s.

Indichiamo inoltre con:

k indice identificativo di unità di rilevazione ($k \in U_L$) ;

Y_k valore assunto dalla caratteristica y oggetto d' indagine nella k-esima unità di rilevazione;

$\underline{X}_k = (X_{1k},..., X_{jk},...,X_{Jk})'$ vettore colonna contenente i valori assunti dalle J variabili ausiliarie $x_1,...,x_j,...,x_J$ nella unità k;

$t_y = \sum_U Y_k$ totale nella popolazione della caratteristica y;

$\underline{t}_x = \left(t_{x_1},...,t_{x_j},...,t_{x_J}\right)'$ vettore colonna contenente i totali noti delle J variabili ausiliarie $x_1,...,x_j,...,x_J$;

π_k probabilità di inclusione dell' unità k ($k \in U_L$);

$D_k=(1/\pi_k)$ peso diretto attribuito alla unità k ($k \in s^*$);

$\underline{D}$ vettore colonna contenente i pesi diretti $(D_1,...,D_k,...,D_n)$ delle n unità campionarie appartenenti all' insieme s;

W_k peso finale attribuito alla unità k ($k \in s$) ;

$\underline{W}$ vettore colonna contenente i pesi finali $(W_1,...,W_k,...,W_n)$ delle n unità campionarie appartenenti all' insieme s;

Nella presente esposizione il simbolo: $\sum_A(\bullet)$ indica, con riferimento ad un generico insieme A, la sommatoria estesa a tutti gli elementi ad esso appartenenti. all'insieme A.

Il nostro obbiettivo è quello di stimare il totale t_y mediante uno stimatore che: (i) sia approssimativamente non distorto rispetto al disegno p(•); (ii) garatisca l'ottenimento di stime dei totali $\underline{t}_x$ che coincidano con i valori noti di tali totali; (iii) attenui l'effetto distorsivo dovuto alla presenza di mancate risposte totali ($n < n^*$); (iv) attenui l'effetto dovuto alla sottocopertura della lista U_L (rispetto alla lista U) da cui è selezionato il campione s*;

Qualora non si verifichi il fenomeno delle mancate risposte ($s^* \equiv s$) e la lista da cui è selezionato il campione non presenti il fenomeno della sottocopertura ($U \subseteq U_L$), lo stimatore diretto del totale t_y, garantisce il rispetto della condizione (i) ed ha la seguente espressione[1]:

$$\tilde{t}_{y\pi} = \sum_s Y_k D_k = \sum_s Y_k / \pi_k. \qquad [1]$$

Tuttavia, nelle situazioni che generalmente si presentano nelle indagini campionarie condotte su larga scala, lo stimatore diretto non garantisce il rispetto delle quatto condizioni sopra elencate; infatti:

- la stima diretta, applicata sul vettore delle J variabili ausiliarie:

$$\underline{\tilde{t}}_{x\pi} = \sum_s \underline{X}_k D_k \qquad [2]$$

[1]Lo stimatore in oggetto è noto in letteratura con il nome di stimatore di Horvitz Thompson (1952).

non coincide con il vettore dei totali noti ($\underline{t}_x$) delle variabili stesse;

- in presenza del fenomeno delle mancate risposte totali o di quello della sottocopertura della lista da cui è selezionato il campione, la stima diretta $\tilde{\underline{t}}_{y\pi}$ si configura come una sottostima del totale t_y, in quanto:

$$E_p\left(\tilde{\underline{t}}_{y\pi}\right) < t_y \ ,$$

in cui $E_p(\bullet)$ rappresenta il valore atteso rispetto al disegno di campionamento $p(\bullet)$.

2.2. La classe degli stimatori di ponderazione vincolata

Una classe di stimatori del totale t_y, che sotto ipotesi piuttosto generali, consente di rispettare le quattro condizioni (i)-(iv) del paragrafo precedente, è quella degli stimatori di ponderazione vincolata[2]. Uno stimatore appartenente a tale classe può essere definito nel modo seguente:

$$\tilde{t}_{yw} = \sum_s Y_k W_k \qquad [3]$$

in cui il vettore dei pesi $\underline{W}$ è ottenuto come soluzione di un sistema di minimo vincolato. La funzione obbiettivo è data da:

$$\min\left[E_p\left\{\sum_s G(D_k, W_k)\right\}\right] \qquad [4]$$

[2]Nella letteratura di lingua anglosassone sull' argomento, gli stimatori in oggetto sono indicati con il termine *calibration estimators* (Deville e S•rndal, 1992).

dove $G(D_k, W_k)$ è una funzione generale di distanza tra il peso diretto D_k ed il peso finale W_k tale che:

(a) per ogni fissato $D_k > 0$, $G(D_k, W_k)$ sia differenziabile rispetto a W_k, strettamente convessa e definita in un intervallo I_{D_K} contenente D_k in cui $G(D_k, D_k) = 0$;

(b) indicando con $g(D_k, W_k) = \dfrac{\delta G(D_k, W_k)}{\delta W_k}$ la derivata prima di $G(D_k, W_k)$, essa deve essere, nell' insieme I_{D_K}, una funzione strettamente crescente di W_k ed inoltre $g(D_k, D_k) = 1$.

I vincoli sono costituiti dal seguente sistema lineare di J equazioni in n incognite[3]: :

$$\sum_s W_k \underline{X}_k = \underline{t}_x \qquad [5]$$

Il nostro obbiettivo è, quindi, quello di individuare un vettore di pesi finali $\underline{W}$ che consenta di rispettare il sistema di vincoli [5] e che contemporaneamente modifichi il meno possibile il vettore dei pesi diretti $\underline{D}$[4].

Il presente paragrafo è finalizzato all'illustrazione della soluzione algebrica del sistema definito dalle relazioni [4] e [5].

[3]Facciamo notare che le condizioni (a) e (b) definiscono le proprietà della funzione di distanza in modo tale che il sistema di minimo vincolato definito dalla •4 • e dalla [5] ammetta una soluzione finita; esse inoltre garantiscono che tale soluzione sia unica .

[4]Il vettore $\underline{W}$ è determinato in modo da essere il più vicino in media , sulla base di una metrica prescelta, al vettore dei pesi diretti $\underline{D}$.

Le proprietà statistiche della classe degli stimatori di ponderazione vincolata saranno descritte dettagliatemente in successivi sccessivi lavori. Riteniamo, tuttavia, utile anticipare qui di seguito le principali proprietà degli stimatori di ponderazione vincolata:

- in assenza dei fenomeni di mancata risposta e di sottocopertura, le stime ottenute utilizzando i pesi finali risultano, per campioni sufficentemente grandi, non distorte, come le stime dirette e più efficienti di quest'ultime, in quanto sostanzialmente basate su una post-stratificazione del campione;
- in presenza dei fenomeni di mancata risposta e di sottocopertura, gli stimatori di ponderazione vincolata consonteno in generale di attenuare i gli effetti distorsivi dovuti alla presenza dei fenomeni in oggetto; inoltre, qualora si verifichino particolari condizioni sui modelli probabilistici che generano la mancata risposta (Binder, 1988) e la sottocopertura (Alexander, 1987) si dimostra che gli stimatori in oggetto conducono a stime corrette.

Passando ora, alla descrizione della soluzione algebrica del problema di minimo vincolato [4] e [5], osserviamo che la risoluzione del problema in oggetto equivale a ricercare nell' insieme delle $\infty^{(n-J)}$ soluzioni possibili del sistema dei vincoli [5], il vettore $\underline{W}$ che soddisfa la [4] minimizzando il valore atteso della somma $\sum_s G(D_k, W_k)$ al variare di s nell'insieme di tutti i campioni possibili. Ciò equivale a minimizzare, per ogni particolare campione estratto s, la quantità $\sum_s G(D_k, W_k)$ nel rispetto dei vincoli [5]. La

[4] può essere, pertanto, sostituita dalla seguente espressione equivalente:

$$\min\left[\sum_s G(D_k, W_k)\right] \qquad [6]$$

che rappresenta il valore minimo della somma $\sum_s G(D_k, W_k)$ condizionato al campione osservato s.

Il vettore $\underline{W}$, soluzione del sistema [5] e [6], si ottiene nel modo qui di seguito illustrato. Si definisce la funzione di Lagrange attraverso l'espressione:

$$L(\underline{\lambda}, \underline{W}) = \sum_s G(D_k, W_k) - \left[\sum_s W_k \, \underline{X}_k - \underline{t}_x\right]\underline{\lambda} \qquad [7]$$

in cui $\underline{\lambda}$ è il vettore di dimensione (J x 1) dei moltiplicatori di Lagrange.

Si risolve, quindi, il sistema omogeneo di (n + J) equazioni :

$$\begin{cases} \dfrac{\delta L(\underline{\lambda}, \underline{W})}{\delta W_k} = g(D_k, W_k) - \underline{X}_k' \, \underline{\lambda} = 0 & \text{per } k \in s \\ \dfrac{\delta L(\underline{\lambda}, \underline{W})}{\delta \lambda_j} = \sum_s X_{jk} \, W_k - t_{xj} = 0 & \text{per } j = 1, \ldots, J. \end{cases} \qquad [8]$$

Se esiste una soluzione del sistema [8], le assunzioni (a) e (b) garantiscono che essa è unica ed è definita da:

$$W_k = D_k \, F\left(\underline{X}'_k \underline{\lambda}\right), \qquad [9]$$

dove $F(\bullet) = g^{-1}(\bullet)$ è la funzione inversa di $g(\bullet)$ avente le seguenti caratteristiche $F(0) = 1$ ed $F'(0) > 0$. Dalla precedente relazione fondamentale si desume che il generico peso W_k del vettore $\underline{W}$ si ottiene moltiplicando il peso diretto corrispondente D_k per un coefficente di correzione scalare $F\left(\underline{X}'_k \underline{\lambda}\right)$, funzione del vettore di varibili ausiliarie $\underline{X}'_k$ e dei J valori incogniti del vettore $\underline{\lambda}$. Al fine di illustrare, i passaggi algebrici che conducono alla relazione [9], consideriamo il caso, valido nella maggior parte delle applicazioni, in cui la funzione $g(W_k , D_k)$ è funzione del singolo argomento W_k / D_k; pertanto, le prime n equazioni del sistema [8] possono essere espresse nel seguente modo:

$$g\,(W_k / D_k) = \underline{X}_k \underline{\lambda}.$$

Introducendo la funzione inversa $g^{-1}(\bullet)$ di $g(\bullet)$, si ha:

$$g^{-1}\left(g\left(W_k / D_k\right)\right) = g^{-1}\left(\underline{X}'_k \, \underline{\lambda}\right);$$

si perviene, quindi all'espressione cercata:

$$W_k = D_k \, g^{-1}\left(\underline{X}'_k \, \underline{\lambda}\right) = D_k \, F\left(\underline{X}'_k \, \underline{\lambda}\right).$$

La [9] non è ancora una relazione operativa nel senso che non permette il calcolo del vettore dei pesi finali $\underline{W}$ in quanto non sono noti i valori numerici del vettore $\underline{\lambda}$. Al fine di pervenire alla

determinazione di $\underline{\lambda}$ moltiplichiamo entrambi i membri della [9] per $\underline{X}_k$, e sommiamo sul campione s; ottenendo in tal modo:

$$\sum_s \underline{X}_k W_k = \sum_s \underline{X}_k D_k \, F\left(\underline{X}'_k \underline{\lambda}\right).$$

Introducendo nell'espressione precedente il sottosistema dei vincoli espressi nelle ultime J equazioni del sistema [8], si ha:

$$\underline{t}_x = \sum_s \underline{X}_k D_k \, F\left(\underline{X}'_k \underline{\lambda}\right)$$

che è equivalente a:

$$\underline{t}_x - \underline{\tilde{t}}_{x\pi} = \sum_s \underline{X}_k D_k \, F\left(\underline{X}'_k \underline{\lambda}\right) - \underline{\tilde{t}}_{x\pi} =$$

$$= \sum_s \underline{X}_k D_k \left(F\left(\underline{X}'_k \underline{\lambda}\right) - 1\right) = \underline{t}_x - \underline{\tilde{t}}_{x\pi}.$$

Infine, indicando con:

$$\underline{\phi}(\underline{\lambda}) = \sum_s \underline{X}_k D_k \left(F\left(\underline{X}'_k \underline{\lambda}\right) - 1\right),$$

si ottiene il seguente sistema di J equazioni nelle J incognite ($\lambda_1,\ldots,\lambda_j,\ldots,\lambda_J$):

$$\underline{\phi}(\underline{\lambda}) = \sum_s \underline{X}_k D_k \left(F\left(\underline{X}'_k \underline{\lambda}\right) - 1\right) = \underline{t}_x - \underline{\tilde{t}}_{x\pi} \qquad [10]$$

dove la generica equazione del sistema è data da:

$$\phi_j(\underline{\lambda}) = \sum_s X_{jk} D_k \left(F\left(\underline{X}'_k \underline{\lambda}\right) - 1\right) = t_{x_j} - \tilde{t}_{x_j\pi}, \qquad [11]$$

essendo:

$$\tilde{t}_{x\,j\pi} = \sum_s X_{jk} D_k$$

Risolvendo il sistema [10] rispetto a $\underline{\lambda}$ si perviene alla soluzione cercata.

Da quanto detto, risulta chiaro che, una volta definita la funzione di distanza $G(D_k,W_k)$ e quindi $F\left(\underline{X}'_k \underline{\lambda}\right)$, è possibile determinare il vettore dei pesi finali $\underline{W}$, introducendo nella [9] la soluzione della [10] rispetto a $\underline{\lambda}$.

Per quanto riguarda la soluzione della [10] rispetto a $\underline{\lambda}$ distinguiamo due casi: il caso in cui $\underline{\phi}(\underline{\lambda})$ sia di tipo lineare e il caso in cui essa sia di tipo non lineare. Nel primo caso $\underline{\phi}(\underline{\lambda})$ può, in generale, essere espressa dalla relazione:

$$\underline{\phi}(\underline{\lambda}) = \underline{T}\ \underline{\lambda} \qquad [12]$$

dove $\underline{T}$ è una matrice simmetrica di (J x J). La soluzione cercata è espressa da:

$$\underline{\lambda} = \underline{T}^{-1}\left(\underline{t}_x - \underline{\tilde{t}}_{x\pi}\right) \qquad [13]$$

Invece, nel caso in cui $\underline{\phi}(\underline{\lambda})$ sia una funzione non lineare di $\underline{\lambda}$, la soluzione del sistema di J equazioni non lineari [10] si ottiene iterativamente attraverso un algoritmo basato sul metodo di Newton. Alla prima iterazione (ν=0) si pone $\underline{\lambda}(\nu=0) = \underline{0}$, dove $\underline{0}$ è un vettore di dimensione (J x 1) i cui elementi sono tutti pari a zero. I valori $\underline{\lambda}(\nu)$ alle succesive iterazioni (ν =1,2,...) sono dati da:

$$\underline{\lambda}(\nu) = \underline{\lambda}(\nu-1) + \left\{\underline{\phi}'\left[\underline{\lambda}(\nu-1)\right]\right\}^{-1}\left\{\underline{t}_x - \underline{\tilde{t}}_{x\pi} - \underline{\phi}\left[\underline{\lambda}(\nu-1)\right]\right\} \qquad [14]$$

in cui: $\underline{\phi}\left[\underline{\lambda}(\nu-1)\right]$ è il vettore i cui J valori sono ottenuti ponendo nella [10] $\underline{\lambda} = \underline{\lambda}(\nu-1)$; e $\underline{\phi}'(\underline{\lambda}(\nu-1))$ è una matrice simmetrica di dimensione (J x J), il cui generico elemento $\phi'_{ji}(\underline{\lambda}(\nu-1))$ sulla riga j-esima e sulla colonna i-esima è la derivata prima di $\phi_j(\underline{\lambda})$ rispetto a λ_i, calcolata ponendo $\underline{\lambda} = \underline{\lambda}$ (ν-1), ossia:

$$\phi'_{ji}\left[\underline{\lambda}(\nu-1)\right] = \left[\frac{\delta\phi_j(\underline{\lambda})}{\delta\lambda_i}\right]_{\underline{\lambda}=\underline{\lambda}((\nu-1))} \qquad [15]$$

L' iterazione finale è quella che verifica almeno una delle seguenti due condizioni:

$$\max_{j=1}^{J}\left(\frac{\left|\lambda_j(\nu-1) - \lambda_j(\nu)\right|}{\left|\lambda_j(\nu-1)\right|}\right) \le C1 \qquad [16]$$

$$\nu = MAXI \qquad [17]$$

dove C1 è una costante scelta nell' intervallo (0 , 1), e MAXI indica il numero massimo di iterazioni ammesse, oltre il quale si giudica che l'algoritmo non converga[5]. Mediante la [16] si interrompe il

[5]Nel software sviluppato sull'argomento, in assenza di ulteriori specificazioni, si pone: $C1=10^{-4}$ e C2 = 100.

processo iterativo quando tra l'iterazione ν e l'iterazione precedente (ν-1) la maggiore differenza relativa sui valori dei λ_j (j=1,...,J) è minore di un valore piccolo a piacere. La condizione [17] viene introdotta al fine di interrompere le iterazioni quando il processo non converge.

2.3. Scelta della funzione di distanza

Consideriamo ora due particolari espressioni della funzione di distanza $G(D_k , W_k)$ che si ritengono utili a risolvere la maggior parte dei problemi di stima che si pongono nelle indagini su larga scala[6]:

$$G(D_k, W_k) = \frac{(D_k - W_k)^2}{Q_k D_k}, \qquad [18]$$

$$G(D_k, W_k) = \frac{W_k}{Q_k} \operatorname{Ln}\left(\frac{W_k}{D_k}\right), \qquad [19]$$

in cui $1/Q_k$ indica un peso non correlato a D_k assegnato all'unità k. Nella maggior parte delle applicazioni si utilizza il peso uniforme $1/Q_k = 1$, ma in alcuni casi può essere conveniente utilizzare pesi $1/Q_k$ variabili[7].

[6]Nel lavoro di Deville e S•rndal (1992), vengono introdotte anche altre funzioni di distanza le quali, però non garantiscono l'esistenza di una soluzione.
[7]Nel lavoro di Alexander (1987), si dimostra l'utilità dell' adozione dei pesi $1/Q_k$ variabili, per risolvere particolari problemi di sottocopertura.

I pesi finali relativi alle funzioni di distanza appena introdotti, sono espressi da:

$$W_k = D_k\left[1 + Q_k \underline{X}_k'\left(\sum_s \underline{X}_k \underline{X}_k' D_k\right)^{-1}\left(\underline{t}_x - \underline{\tilde{t}}_{x\pi}\right)\right], \qquad [20]$$

per quanto riguarda la [18]; e da:

$$W_k = D_k \exp\left(Q_k \underline{X}_k' \underline{\lambda}\right), \qquad [21]$$

per quanto riguarda la [19]. In quest'ultima espressione, valori di $\underline{\lambda}$ vengono calcolati in modo iterativo attraverso la relazione [14] in cui:

$$\underline{\phi}(\underline{\lambda}(\nu-1)) = \sum_s \underline{X}_k D_k \left\{\exp\left(Q_k \underline{X}_k' \underline{\lambda}(\nu-1)\right) - 1\right\},$$

$$\underline{\phi}'(\underline{\lambda}(\nu-1)) = \sum_s \underline{X}_k \underline{X}_k' D_k \exp\left(Q_k \underline{X}_k' \underline{\lambda}(\nu-1)\right)$$

Esaminiamo ora come sono state ottenute le precedenti relazioni [20] e [21]. Per la prima] si è ricavata la funzione $F\left(\underline{X}_k' \underline{\lambda}\right)$ relativa alla funzione di distanza [18]; sostituendo poi nella [9] il valore esplicito di $\underline{\lambda}$, espresso mediante una relazione analoga alla [13], si è pervenuti alla [20]. La [21], invece, è stata ottenuta

ricavando la funzione $F(\underline{X}_k' \underline{\lambda})$ relativa alla [18] e sostituendo poi tale espressione nella [9].

Le funzioni di distanza qui introdotte hanno la desiderabile proprietà di portare sempre ad una soluzione qualora il sistema dei vincoli sia congruente.

Si dimostra (Deville and Särndal, 1992, p.p 379) che, per campioni sufficentemente grandi, la [18] e [19] conducono a stimatori aventi approssimativamente la stessa varianza; pertanto al fine di pervenire ad una scelta tra le due funzioni di distanza introdotte è necessario analizzare l'intervallo dei valori che i coefficienti di correzione $F(\underline{X}_k' \underline{\lambda})$ assumono nei due casi.

La [18] è la funzione di distanza che conduce allo stimatore di regressione generalizzato (Särndal, Swensson e Wretman, 1992; Isaki and Fuller, 1982). Lo stimatore in oggetto viene adottato per l'ottenimento delle stime dell'indagine canadese sulle Forze di Lavoro ed è stata applicato in ambito ISTAT per il calcolo delle stime dell'indagine Sulle Condizioni di Salute della Popolazione e sul Ricorso ai Servizi Sanitari 1986-1987 (ISTAT , 1991). Essa porta a coefficenti di correzione che possono variare nell'intervallo $(-\infty , \infty)$ e quindi condurre anche a pesi W_k negativi, i quali potrebbero essere non accettabili in alcune applicazioni.

La [19] è la funzione di distanza che viene utilizzata per l'ottenimento delle stime di massima verosimiglianza dei modelli log-lineari (Darroch and Ratcliff, 1972) ed è stata adottata in ambito ISTAT per il calcolo dei pesi finali dell'indagine Multiscopo sulle Famiglie (ISTAT ,1993). Essa porta a coefficenti di correzione che

possono variare nell'intervallo $(0 , \infty)$ e conduce quindi a pesi finali W_k sempre positivi. Tuttavia, in alcuni casi non favorevoli, i pesi finali possono presentare valori estremamente grandi rispetto ai corrispondenti pesi base D_k, , risultando pertanto non accettabili . Infatti, applicando questi pesi per l'ottenimento di stime per varie sottopopolazioni in differenti domini di studio è possibile ottenere stime non realistiche relativamente ad alcuni domini.

2.4. Funzione generalizzata di distanza

Da quanto detto nel paragrafo precedente risulta evidente che i pesi finali ottenuti attraverso le funzioni di distanza [18] e [19] possono presentare problemi in particolari ambiti applicativi. Qui di seguito presentiamo una funzione di distanza che permette di superare tali problemi, in quanto porta a pesi finali aventi le seguenti caratteristiche:

- i valori possibili dei coefficenti di correzione dei pesi diretti sono limitati in un intervallo definito a priori;
- i pesi finali espressi dalle relazioni [20] e [21] si ottengono come casi particolari definendo in modo opportuno l'intervallo di correzione dei pesi diretti.

La funzione di in oggetto è data da: $G(D_k, W_k) = \left(\frac{W_k}{D_k} - L\right) \ln \frac{\frac{W_k}{D_k} - L}{1 - L} + \left(U - \frac{W_k}{D_k}\right) \ln \frac{U - \frac{W_k}{D_k}}{U - 1}$, [22]

dove: L ed U sono due costanti, da scegliere in modo opportuno, tali che: L<1<U.

Ricavando la funzione $F(\underline{X}_k' \underline{\lambda})$ e sostituendo la sua espressione nella [9] si determina l'espressione del generico peso finale in funzione del vettore incognito $\underline{\lambda}$:

$$W_k = D_k F(\underline{X}_k' \underline{\lambda}) = D_k \frac{L(U-1) + U(1-L)\exp(A\, Q_k\, \underline{X}_k'\, \underline{\lambda})}{(U-1) + (1-L)\exp(A\, Q_k\, \underline{X}_k'\, \underline{\lambda})}, \quad [23]$$

essendo A è una costante definita da:

$$A = \frac{U-L}{(1-L)(U-1)}. \quad [24]$$

I valori di $\underline{\lambda}$ vengono calcolati in modo iterativo attraverso la relazione [14]; in cui:

$$\underline{\phi}(\underline{\lambda}(\nu-1)) = \sum_s \underline{X}_k\, D_k \left\{ \frac{L(U-1) + U(1-L)\exp(A\, Q_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))}{(U-1) + (1-L)\exp(A\, Q_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))} - 1 \right\} \quad [25]$$

$$\phi'_{ji}(\underline{\lambda}(\nu-1)) = \sum_s X_{jk}\, X_{ik}\, D_k \frac{(U-L)^2 \exp(AQ_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))}{\left[(U-1) + (1-L)\exp(AQ_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))\right]^2}$$

. [26]

Il generico elemento j (j=1,...,J) del vettore $\underline{\phi}(\underline{\lambda}(\nu-1))$ è dato da:

$$\phi_j(\underline{\lambda}(\nu-1)) = \sum_s X_{jk}\, D_k \left\{ \frac{L(U-1) + U(1-L)\exp(AQ_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))}{(U-1) + (1-L)\exp(AQ_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))} - 1 \right\}$$

inotre, il generico elemento $\phi'_{ji}(\underline{\lambda}(\nu-1))$ sulla riga j-esima e sulla colonna i-esima della matrice $\underline{\phi}'(\underline{\lambda}(\nu-1))$ è definito da:

$$\phi'_{ji}(\underline{\lambda}(\nu-1)) = \sum_s X_{jk}\, X_{ik}\, D_k \frac{(U-L)^2 \exp(AQ_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))}{\left[(U-1) + (1-L)\exp(AQ_k\, \underline{X}_k'\, \underline{\lambda}(\nu-1))\right]^2}$$

La [22] è una funzione di distanza che conduce a pesi finali W_k compresi nell' intervallo $LD_k \le W_k \le UD_k$. Ponendo $L \ge 0$, i pesi sono sempre positivi. Scegliendo un valore di L negativo e molto grande in valore assoluto ed un valore di U molto grande (ad esempio, L=-1.000, U=1000), la soluzione trovata approssima quella data dalla [18]; con un valore di L positivo e molto piccolo ed un valore di U molto grande (ad esempio L=0,0001, U=1000) si approssima la soluzione data dalla [19].

3. Procedura SAS per il calcolo dei pesi finali

3.1. Descrizione generale

Qui di seguito descriveremo le principali caratteristiche di una procedura informatica scritta in linguaggio SAS che permette di calcolare i pesi finali mediante lo stimatore di ponderazione vincolata presentato nel precedente paragrafo 2.4.

Questa procedura è stata creata dagli autori ed è disponibile presso l'Istat.

Gli elementi fondamentali della procedura sono:

- **generalità di applicazione:** può essere applicata alla maggiorparte delle indagini campionarie ISTAT sulle famiglie e sulle imprese;
- **possibilità di definire differenti domini di studio**: per dominio di studio si intende un particolare sottoinsieme della popolazione rispetto al quale si vogliono calcolare le stime oggetto di indagine e si dispone di totali noti su cui vincolare le stime stesse. Ad esempio, nelle indagini sulle famiglie, un dominio può essere individuato dalla popolazione residente di una determinata regione ed i totali noti possono essere dati dalla struttura della popolazione per sesso e classi di età; nelle indagini sulle imprese, un dominio può essere identificato dall' insieme delle imprese di una determinata attività economica;
- **facilità di utilizzazione**: può essere utilizzata direttamente dall'utente che è responsabile dell'indagine, poichè l'applicazione della procedura per le singole indagini campionarie dell'Istituto

richiede solo poche e semplici modifiche ai programmi di cui essa è composta;

- **parametrizzazione:** la procedura dipende da un'insieme di parametri che l'utente deve specificare. Al variare di tali parametri si modifica il vettore dei pesi finali che si ottiene;
- **ricchezza di informazioni statistiche sul vettore dei pesi finali ottenuto:** per ciascun dominio di studio (dominio di pubblicazione delle stime) , viene prodotto come output della procedura un'insieme di statistiche sul vettore dei pesi diretti $\underline{D}$ il vettore dei correttori $\underline{F}$ e quello dei pesi finali $\underline{W}$. Dall'analisi di tali statistiche è possibile scegliere tra le soluzioni alternative ottenute, per il vettore dei pesi finali $\underline{W}$, al variare di differenti specificazioni dei parametri della procedura.

La procedura richiede come input i data set SAS: **NOTI.TOTALI, STIME.TOTALI** e **DATI.CAMPIONE** ed alcuni parametri che l'utente ha la possibilità di definire con valori differenti da quelli già definiti per default. Essa produce come output alcuni data set SAS ed alcuni file CMS; i data set SAS sono: **PESI.PESI, STAT.STAT, STIME.STIME**; i file CMS sono **PESI DATA , STIMA1 LISTING** e **STIMA4 LISTING.**

La procedura che gira in ambiente CMS è composta dai 4 programmi SAS: **STIMA1, STIMA2, STIMA3** e **STIMA4**. Essa viene lanciata digitando sul terminale il comando: **STIMA** e premendo successivamente il tasto di **INVIO**. In particolare, le azioni svolte da ciascuno dei 4 programmi appena menzionati, sono le seguenti:

- **STIMA1**: legge i data set SAS di input, crea per ciascuno di essi la variabile CONTA (che assegna un numero d'ordine progressivo ai diversi domini di studio) e produce il file CMS STIMA1 LISTING contenente alcune stampe di controllo sui file di input letti;
- **STIMA2**: legge i file SAS di input, calcola i pesi finali W_k (mediante la procedura iterativa descritta nel paragrafo 2.4) con riferimento al primo dominio di studio (CONTA=1) e crea i file di output PESI.PESI, STIME.STIME e STAT.STAT con le infomazioni relative al primo dominio di studio;
- **STIMA3**: legge i files SAS di input, calcola i pesi finali W_k (mediante la procedura iterativa descritta nel paragrafo 2.4) con riferimento ai domini di studio successivi al primo (CONTA>1) e aggiorna i file di output PESI.PESI, STIME.STIME e STAT.STAT con le informazioni relative a tali domini;
- **STIMA4**: legge i files SAS di output, crea il file CMS PESI DATA dei pesi finali W_k e produce il file CMS STIMA4 LISTING contenente le stampe di controllo sui data set SAS creati come output.

La procedura informatica appena illustrata è generale, nel senso che può essere applicata per calcolare i pesi finali delle differenti indagini campionarie dell'Istat senza operare nessuna modifica ai programmi SAS di cui è composta. E' richiesta unicamente la creazione da parte dell' utente dei 3 data set SAS di input prima specificati la cui struttura è descritta in dettaglio nel paragrafo 3.3. Inoltre l'utente ha la possibilità di definire per ciascun dominio di studio valori differenti da quelli già fissati per default per i parametri

L ed U (di cui al paragrafo 2.4) che limitano la variabilità dei pesi finali nell'intervallo $LD_k \leq W_k \leq UD_k$ e per il numero massimo di iterazioni ammesse MAXI definito dalla formula [17]. La modifica dei valori di default dei parametri L, U e MAXI richiede da parte dell' utente un intervento, sui programmi STIMA2 e STIMA3 per cambiare i valori assegnati alle variabili MACRO rispettivamente indicate con **&L**, **&U** ed **&MAXI**. Le modalità con cui effettuare l'intervento in questione saranno descritte in dettaglio nel paragrafo 3.4.

La procedura non richiede alcuna modifica nel caso in cui il numero dei domini oggetto di studio è minore o uguale a 100; altrimenti per un numero maggiore di 100 è necessario aggiungere alcune istruzioni al programma STIMA3 SAS, secondo le modalità descritte nel paragrafo 3.5.. Le istruzioni per l'esecuzione della procedura sono illustrate nel paragrafo 3.2.

In conclusione facciamo presente che: nel paragrafo 3.6. sono illustrati gli output della procedura; nel paragrafo 3.7. sono approfondite le modalità di applicazione della procedura nelle indagini concrete: il paragrafo 3.7.2. è relativo alle indagini sulle famiglie ed il paragrafo 3.7.3 si riferisce alle indagini sulle imprese.

3.2. Comandi per l'esecuzione della Procedura

In tutti i casi in cui il numero dei domini di studio è minore od uguale a 100 e non si intende modificare i valori di default dei parametri L, U e MAXI, la procedura viene eseguita digitando sul

terminale il comando: **STIMA** e premendo successivamente il tasto di **INVIO**. Tale operazione, richiede naturalmente che si siano formati i data set di input descritti nel paragrafo 3.3.

Nel caso in cui il numero dei domini di studio sia maggiore di 100, l' utente deve aggiungere alcune istruzioni alla fine del programma STIMA3 SAS, come descritto nel paragrafo 3.5. [8]. Successivamente, la procedua può essere lanciata digitando sul terminale il comando STIMA.

E' opportuno che l'utente faccia girare una prima volta la procedura utilizzando i valori di default dei parametri L, U e MAXI. In seguito, se dall' analisi dei tabulati prodotti dalla procedura, si desume la necessità, per alcuni domini di studio, di modificare i valori dei parametri in parola l'utente deve intervenire nei programmi STIMA2 SAS e STIMA3 SAS, secondo le modalità descritte nel paragrafo 3.4., e successivamente lanciare la procedura digitando il comando STIMA.

3.3. Data set di input della procedura

I data set SAS di input sono :NOTI.TOTALI, STIME.TOTALI e DATI.CAMPIONE. Essi contengono tutte variabili di tipo numerico devono essere formati secondo le modalità di seguito descritte.

[8]Per una piena compresione del paragrafo 3.5. è opportuno che il lettore consulti prima il paragrafo 3.4. .

NOTI.TOTALI, è il data set SAS che contiene i valori dei J totali noti t_X con riferimento a ciascun dominio di studio. Pertanto, se indichiamo con A il numero complessivo dei domini di studio prescelti (ad esempio nel caso in cui i domini sono le regioni, A è pari a 21) , il data set in analisi è formato da A osservazioni e **J+1** variabili. Le J+1 variabili sono indicate nell'ordine con i nomi : DOMINIO, TX1, TX2,...,TXJ e contengono rispettivamente, per ciascuno degli A domini di studio, il codice del dominio ed i J valori dei totali noti per tale dominio. Ad esempio se i domini di studio sono le regioni ed il numero dei totali noti è J=10 si ha che il data set contiene 21 osservazioni (una per regione) ed 11 variabili: DOMINIO, TX1-TX10; dove la variabile DOMINIO riporta il codice di regione dell' osservazione.

STIME.TOTALI, è il data set SAS che contiene i valori delle stime dirette, $\tilde{t}_{x\pi}$, dei J totali noti t_X con riferimento a ciascun dominio di studio. Esso è formato da A osservazioni e **J+1** variabili. Le J+1 variabili sono indicate nell'ordine con i nomi: DOMINIO, SX1, SX2,...,SXJ e contengono rispettivamente, per ciascuno degli A domini di studio, il codice del dominio ed i J valori delle stime dirette dei corrispondenti totali noti per tale dominio. Con riferimento all'esempio sopra riportato in cui J=10 si ha che il data set contiene 21 osservazioni ed 11 variabili: DOMINIO, SX1-SX10.

DATI.CAMPIONE, è il data set SAS contenente i valori che le J variabili ausiliarie X_k assumono con riferimento alle unità campionarie rispondenti di ciascuno degli A domini di studio.

Pertanto se indichiamo con n il numero totale di unità campionarie rispondenti all'indagine, il data set in analisi è composto di **n** osservazioni e **J+3** variabili. Le J+3 variabili sono indicate nell'ordine con i nomi CODICE, COEF, DOMINIO, X1, X2,...,XJ e contengono rispettivamente, per ciascuna delle n unità campionarie, il codice, k , identificativo dell'unità, il peso diretto, D_k , il codice del dominio di studio cui l'unità appartiene ed i J valori che le variabili ausiliarie assumono per tale unità. Con riferimento all'esempio sopra riportato in cui J=10 si ha che il data set contiene le 13 variabili: CODICE, COEF, DOMINIO, X1-X10.

3.4. Modifica dei valori di default dei parametri L , U *e* MAXI.

Ricordiamo che i parametri L e U, definiti nella formula [22], limitano la variabilità ammessa per i pesi finali nell'intervallo $LD_k \leq W_k \leq UD_k$; il parametro MAXI, definito dalla formula [17], rappresenta il numero massimo di iterazioni ammesso per l'individuazione della soluzione finale. La procedura utilizza i seguenti valori di default per i parametri in oggetto: L=0, U=1.000, MAXI=15. L' utente può modificare i parametri L, U e MAXI cambiando i valori rispettivamente attribuiti alle variabili MACRO &L, &U e &MAXI. Tale modifica deve essere effettuata per ciascuno dei domini oggetto di studio.

Prima di descrivere le azioni da intraprendere per modificare i valori di default dei parametri L, U e MAXI è utile precisare che la

procedura utilizza 2 variabili (DOMINIO e CONTA) per codificare i domini oggetto di studio:

- la variabile DOMINIO (cfr. par 3.3.) è quella definita dall' utente che crea i data set di input;
- la variabile CONTA viene creata dalla procedura assegnando un numero progressivo ai codici della variabile DOMINIO.

Ad esempio nel caso in cui i domini oggetto di studio siano costituiti dalle regioni italiane comprese le provincie autonome di Bolzano e Trento codificate dall' utente rispettivamente con i codici 41 e 42 si ha la seguente situazione:

Regioni	DOMINIO	CONTA
Pie:monte	1	1
Valle d'Aosta	2	2
Lombardia	3	3
Veneto	5	4
Friuli	6	5
Liguria	7	6
Emilia	8	7
Toscana	9	8
Umbria	10	9
Marche	11	10
Lazio	12	11
Abruzzi	13	12
Molise	14	13

Regioni	DOMINIO	CONTA
Campania	15	14
Puglia	16	15
Basilicata	17	16
Calabria	18	17
Sicilia	19	18
Sardegna	20	19
Bolzano	41	20
Trento	42	21

L' utente per conoscere in modo esatto la codifica progressiva ai domini di studio della variabile CONTA deve lanciare la prima volta la procedura senza apportare modifiche ai programmi, e consultare il file CMS STIMA4 LISTING, in cui nelle prime 2 colonne delle tabelle 1, 2,3 e 4 vengono riportati rispettivamente il codice progressivo di dominio (variabile CONTA) ed il codice di dominio (variabile DOMINIO).

Ciò detto, per modificare i valori di default dei parametri L, U e MAXI si deve agire nel seguente modo:

1. per il **primo dominio**, che presenta un valore pari ad 1 della variabile CONTA, si modificano i valori di L, U e MAXI nell'ultima istruzione del programma **STIMA2 SAS** che, con riferimento ai valori di default assume la forma:

   %MM (1, 0, 1000, 15);

   in cui: il primo numero in parentesi, che non deve essere modificato, indica il valore assunto dalla variabile CONTA; il

secondo, il terzo ed il quarto numero in parentesi (0, 1000, 15) indicano rispettivamente i valori assegnati ai parametri L, U e MAXI. Per modificare tali parametri è necessario cambiare i valori ad essi attribuiti. Ad Esempio, nel caso in cui l'utente voglia definire i seguenti parametri L= - 1.000, U = 30, MAXI = 18, l'ultima istruzione del programma STIMA2 SAS deve essere modificata nel modo seguente:

%MM (1, -1000, 30, 18);

2. per i **domini successivi** l'utente trova alla fine del programma **STIMA3 SAS** le seguenti 99 istruzioni:

   %MM (2, 0, 1000, 15);
   %MM (3, 0, 1000, 15);
   %MM (4, 0, 1000, 15);
   .
   .
   .
   %MM (100, 0, 1000, 15);

   ciascuna delle quali è relativa ad un dominio di studio. Il primo numero tra parentesi, che non deve essere modificato dall'utente, indica il valore del codice progressivo di dominio di studio assunto dalla variabile CONTA, il secondo, il terzo ed il quarto numero in parentesi (0, 1000, 15) indicano rispettivamente i valori assegnati ai parametri L, U e MAXI per i domini di studio identificati dalla variabile CONTA. L' utente deve modificare i valori di L, U e MAXI delle istruzioni %MM corrispondenti ai domini di studio sui quali vuole intervenire. Per riferirci ad un caso concreto, riprendiamo l'esempio precedente dei domini di

studio regionali e supponiamo che l' utente voglia assegnare ai parametri in oggetto i valori di -50, 50 e 30 alle regioni Valle d'Aosta (CONTA=2), Bolzano (CONTA=20) e Calabria (CONTA=17). In tale situazione, le istruzioni %MM alla fine del programma STIMA3 SAS devono essere modificate nel modo seguente:

```
%MM (2, -50, 50, 30);
%MM (3, 0, 1000, 15);
%MM (4, 0, 1000, 15);
%MM (5, 0, 1000, 15);
%MM (6, 0, 1000, 15);
%MM (7, 0, 1000, 15);
%MM (8, 0, 1000, 15);
%MM (9, 0, 1000, 15);
%MM (10, 0, 1000, 15);
%MM (11, 0, 1000, 15);
%MM (12, 0, 1000, 15);
%MM (13, 0, 1000, 15);
%MM (14, 0, 1000, 15);
%MM (15, 0, 1000, 15);
%MM (16, 0, 1000, 15);
%MM (17, -50, 50, 30);
%MM (18, 0, 1000, 15);
%MM (19, 0, 1000, 15);
%MM (20, -50, 50, 30);
%MM (21, 0, 1000, 15);
%MM (22, 0, 1000, 15);
```



```
%MM (100, 1000, 15); .
```

Prima di concludere il paragrafo diamo alcune indicazioni operative, per la scelta dei valori dei parametri L, U e MAXI.

L deve assumere un valore sempre inferiore ad 1, mentre U deve sempre assumere un valore superiore ad 1. In generale sarebbe desiderabile che tali variabili siano definite con valori il più possibile prossimi ad 1. Per ottenere tale risultato si può applicare più volte la procedura iterativa, secondo la modalità di seguito descritta:

a) la prima volta, si applica la procedura assegnando i valori di default L=0, U=1.000;
b) se la procedura iterativa non converge ad una soluzione, o converge solo estendendo il campo di variabilità ammesso per i coefficenti di correzione per valori negativi delparametro L, ciò significa che i J totali noti, t_X , non permettono di definire un sistema congruente di vincoli. In tal caso è necessario pertanto controllare la congruenza dei vincoli e ridune il numero;
c) se, invece, la procedura iterativa converge, si osserva sulla tabella 2 del file CMS di output STIMA4 LISTING, il valore minimo e massimo dei coefficenti di correzione ottenuti per ciascuno degli A domini;
d) si applica nuovamente la procedura definendo, per ciascuno degli A domini di studio, un valore di L maggiore del minimo osservato ed un valore U minore del massimo osservato;

e) si applicano più volte i precedenti punti b) e c) in modo da ridurre progressivamente il campo di variazione dei coefficenti di correzione ottenuti;

f) per ciascuno degli A domini di studio, si interrompe l'applicazione della procedura, quando per i valori assegnati ad L ed U, la procedura non converge più ad una soluzione accettabile;

g) per ciascuno degli A domini di studio, la scelta tra soluzioni equivalenti, in termini di ampiezza dell'intervallo [L , U], può essere fatta mediante la consultazione delle tabelle 1-5 del file CMS di output STIMA4 LISTING. La soluzione ottimale deve avere, comunque, le seguenti caratteristiche: i) valore minimo del coefficente di variazione dei correttori; ii) minimo intervallo di variazione [m , M] tra i valori minimo (m) e massimo (M) ottenuti per i coefficenti di correzione; iii) coefficente di variazione dei pesi finali $\underline{W}$, vicino al coefficente di variazione dei pesi diretti $\underline{D}$; iv) minima differenza in media tra vettore dei totali noti t_X e le corrispondenti stime ottenute mediante i pesi finali $\underline{W}$.

Per quanto riguarda il parametro MAXI, che indica il massimo numero ammesso di iterazioni, in base alle prove effettuate si è notato che quando il sistema dei vincoli è congruente la procedura in genere converge con un numero di iterazioni inferiore a 15. Quando, invece, il sistema dei vincoli non è congruente, un aumento del numero delle iterazioni non porta comunque a soluzione. Se comunque, l' utente accetta soluzioni approssimate, allo scopo di

risparmiare tempo di elaborazione può definire un valore del parametro MAXI inferiore a 15.

3.5. Caso di più di 100 domini di studio

Nel caso in questione l'utente deve aggiungere alla fine del programma STIMA3 SAS tante istruzioni del tipo %MM, prima descritte, una per ciascuno dei domini di studio che hanno la variabile CONTA superiore a 100. Ad esempio nel caso in cui vi siano 105 domini di studio l'utente deve aggiungere alla fine del programma STIMA3 SAS, le seguenti 5 istruzioni:

```
%MM (101, 0, 1000, 15);
%MM (102, 0, 1000, 15);
%MM (103, 0, 1000, 15);
%MM (104, 0, 1000, 15);
%MM (105, 0, 1000, 15);.
```

Come si nota dall'esempio l'utente deve avere cura di codificare, nel primo numero tra parentesi il codice progressivo di dominio espresso dalla variabile CONTA.

3.6. Output della procedura

Come già detto la procedura produce come output dei data set SAS e dei file CMS.

I data set SAS di output hanno i seguenti nomi: STAT.STAT, STIME.STIME e PESI.PESI. Essi contengono tutte variabili numeriche e sono formati nel modo di seguito illustrato.

STAT.STAT è il data set SAS contenente i valori delle statistiche sui pesi diretti, sui coefficenti di correzione dei pesi diretti e sui pesi finali per ciascuno degli A domini di studio. Esso, inoltre, contiene per ogni dominio alcune informazioni sull'andamento della procedura iterativa. Il data set in analisi è formato da A osservazioni e 25 variabili. In particolare il data set contiene le seguenti variabili: CONTA, che numera progressivamente i domini di studio; L, U e MAXITER , che denotano rispettivamente i valori assegnati dall' utente ai parametri L, U e MAXI; MAXD, MAXF MAXW, MIND, MINF MINW, SUMD, SUMW, SUMF, VARD, VARW, VARF, MEAND, MEANW, MEANF, CVD, CVW, CVF, indicanti rispettivamente il valore massimo, il valore minimo, la somma, la varianza, la media ed il coefficente di variazione dei pesi diretti (ultima lettera D), dei pesi finali (ultima lettera W) e dei coefficenti di correzione (ultima lettera F); ITER, che è il numero di iterazioni realizzato; R2 che indica il numero di unità campionarie nel dominio; C2 che è il numero J dei vincoli definiti dall'utente.

PESI.PESI è il data set SAS contenente il peso finale W_k , il peso diretto D_k ed il coefficente di correzione F_k per ciascuna delle n unità rispondenti all'indagine. Esso è composto di n osservazioni e 6 variabili indicate nell'ordine con i nomi CODICE, CONTA, DOMINIO, D, F e W che rappresentano rispettivamente, per ciascuna delle n unità campionarie: il codice, k (identificativo dell'unità), il numero progressivo associato a ciascun dominio; il

codice di dominio di studio cui l'unità appartiene, il peso diretto D_k , il coefficente di correzione F_k ed il peso finale W_k.

STIME.STIME è il data set SAS contenente i valori dei totali noti t_x e delle corrispondenti stime campionarie dirette e finali. Esso è composto da (A×J) osservazioni e 8 variabili . Ciascuna osservazione è relativa alla specifica concatenazione di un dominio di studio con la generica variabile x_j. Le variabili del data set sono indicate con i nomi: CONTA, TOTALE, TX, TXD, TXW, DIFFD, DIFFW, G, che denotano, rispettivamente: il numero progressivo associato a ciascun dominio di studio (CONTA), il codice j relativo alla variabile x_j (TOTALE), il totale noto t_{x_j} (TX), la stima diretta (TXD) e la stima finale (TXW) del totale noto t_{x_j}, la differenza assoluta tra totale noto e corrispondente stima diretta (DIFFD), la differenza assoluta tra totale noto e corrispondente stima finale (DIFFW), il valore del moltiplicatore di Lagrange (G).

I file CMS di output hanno i seguenti nomi: STIMA1 LISTING, STIMA4 LISTING e PESI DATA e sono così formati come di seguito illustrato.

STIMA1 LISTING è il file CMS contente le stampe di controllo sui data set SAS di input. In particolare, il file contiene: la stampa delle PROC CONTENTS SAS dei tre data set SAS di input; la stampa del data set NOTI.TOTALI con riferimento alle variabili CONTA, DOMINIO, TX1 e TX2; la stampa del data set STIME.TOTALI con riferimento alle variabili CONTA, DOMINIO, SX1 e SX2. L'utilità di tale file è quella di permettere i controlli di

conformità dei data set SAS di input agli standard definiti nel paragrafo 3.3. In particolare è possibile controllare la conformità dei data set per quanto riguarda le loro dimensioni, i nomi delle variabili in esse contenuti ed infine l'entità delle differenze tra i primi due totali noti e le corrispondenti stime dirette.

STIMA4 LISTING è il data set contenente la stampa di 5 tabelle statistiche sui principali risultati che si ottengono mediante la procedura utilizzata. Le tabelle sono costruite sulla base delle informazioni riportate nei due data set SAS di output : STAT.STAT e STIME.STIME. In particolare, le tabelle 1 2 3 e 4, prodotte in base alle informazioni del data set STAT.STAT sono così formate: la tabella 1 2 e 3, riportano, rispettivamente per ciascuno degli A domini di studio le statistiche sui pesi finali, sui coefficenti di correzione e sui pesi diretti; la tabella 4, invece, riporta, per ciascun dominio, informazioni relative all'andamento della procedura iterativa. La tabella 5, prodotta sulla base delle informazioni desunte dal data set STIME.STIME, riporta le informazioni sui termini noti e le corrispondenti stime dirette e finali in ciascun dominio;

PESI DATA è il file CMS dei pesi finali. Esso è pertanto composta da **n** rekord ciascuno dei quali ha i seguenti tre campi: codice di dominio cui l'unità appartiene, codice identificativo dell'unità e peso finale. Il file è strutturato nel modo seguente:

Posizioni di inizio e fine campo	Contenuto del campo
1-3	Codice di dominio oggetto di studio, (variabile SAS DOMINIO)
5-14	Codice identificativo dell' unità (codice k)
16-27	Peso finale (W_k), le posizioni 26 e 27 sono i primi due decimali del peso.

L' utente, al fine di produrre le tabelle di pubblicazione deve accoppiare, il file PESI DATA al file contenente gli n record con le variabili rilevate nell' indagine campionaria. L'accoppiamento deve avvenire sulla base del codice identificativo dell' unità. Le stime vengono, infine ottenute sulla base della formula [3].

3.7. Modalità di applicazione della procedura alle indagini campionarie ISTAT

3.7.1. Premessa

Il presente paragrafo è finalizzato a fornire i criteri guida relativi alla scelta ed alla definizione dei J totali noti t_x (che costituiscono i vincoli della procedura) e delle corrispondenti J variabili ausiliarie $\underline{X}_k$ da attribuire a ciascuna delle n unità campionarie rispondenti.

Inoltre, sarà descritto il metodo per la costruzione dei pesi base per le indagini ISTAT sulle famiglie (par 3.7.2.) e sulle imprese (3.7.3.).

Si possono sostanzialmente individuare due situazioni informative:

- situazione 1, in cui risultano noti i totali di alcune variabili calcolati su tutta la popolazione oggetto d'indagine (insieme U, par. 2.) e *relativi al periodo di riferimento dell' indagine*. Tali totali possono essere anche conosciuti per particolari sottopopolazioni (che definiscono una partiziona dell'insieme U) che individuano i domini oggetto di studio. Il caso in oggetto è quello tipico delle indagini ISTAT sulle famiglie in cui risulta conosciuto, per ciascuna regione, l'ammontare totale della popolazione secondo le modalità incrociate del sesso e della classe di età [9].
- situazione 2, in cui non vi sono informazioni sulla popolazione oggetto d'indagine aggiornate al periodo di riferimento dell' indagine; mentre risultano noti i totali di alcune variabili desunte dall' archivio da cui è stato selezionato il campione (insieme U_L, par. 2.). Pertanto in questa situazione le informazioni note si riferiscono al *periodo di formazione dell' archivio*. La situazione in oggetto è quella tipica delle indagini ISTAT sulle imprese (o sulle aziende agricole) in cui ad esempio risulta nota la classe di attività economica ed il numero di addetti di ciascuna delle imprese presenti nell' *archivio delle imprese*. Pertanto risultano noti, alla data di formazione dell' archivio, il numero totale delle imprese e l'ammontare totale degli addetti secondo le modalità incrociate della classe di attività economica e della classe di addetti.

[9] I totali in oggetto sono determinati in base a modelli demografici sui dati desunti dalle statistiche demografiche.

Descriviamo ora, in generale, le azioni da intraprendere nella situazione 1:

(i) si individua l'insieme V delle variabili oggetto di rilevazione i cui totali, riferiti all'insieme U delle unità della popolazione oggetto di indagine, risultano noti alla data di riferimento dell'indagine;

(ii) nell'insieme delle variabili di cui al punto (i) , si cerca quel sottoinsieme, v ($v \subseteq V$), di variabili che risultano maggiormente correlate con l'insieme delle variabili oggetto indagine;

(iii) per la definizione del vettore delle J variabili ausiliarie $\underline{X}_k$ e di quello dei J termini noti corrispondenti, si utilizza l'insieme v. Facciamo notare che se si utilizza l'insieme v, il vettore dei pesi finali $\underline{W}$ ottenuto mediante la procedura iterativa ha la proprietà di correggere le stime dirette in presenza del fenomeno della sottocopertura.

Descriviamo, ora, le azioni da intraprendere nella situazione 2:

(i) nell'insieme V_L delle variabili presenti nell'archivio da cui si è selezionato il campione, si cerca quel sottoinsieme v_L ($v_L \subseteq V_L$) di variabili, che risultano maggiormente correlate con l'insieme delle variabili oggetto indagine. Sulla base delle varibili in oggetto è possibile calcolare i totali noti sommando i valori delle variabili sulle unità dell' archivio ;

(ii) per la definizione del vettore delle J variabili ausiliarie $\underline{X}_k$ e di quello dei J termini noti corrispondenti, si utilizza l'insieme v_L. Facciamo notare che se si utilizza l'insieme v_L, il vettore dei pesi

finali $\underline{W}$ ottenuto mediante la procedura iterativa ha la proprietà di correggere le stime dirette in presenza del fenomeno della mancata risposta.

Nella situazione congiunta in cui risultano noti sia i totali di alcune variabili calcolati su tutta la popolazione oggetto d'indagine e relativi al periodo di riferimento dell' indagine (insieme v), sia i totali di alcune variabili desunte dall' archivio da cui è stato selezionato il campione e relativi al periodo di formazione dell'archivio (insieme v_L) si possono utilizzare entrambi gli insiemi per l'ottenimento del vettore $\underline{W}$. In tale situazione, il vettore dei pesi finali $\underline{W}$ ottenuto mediante la procedura iterativa ha la proprietà di correggere le stime dirette nel caso di presenza congiunta dei fenomeni delle mancate risposte totali e della sottocopertura. Il vettore dei pesi finali deve essere ottenuto attraverso due applicazioni successive della procedura iterativa. La prima volta, si utilizza l'insieme di variabili v_L ottenendo un primo vettore dei pesi $\underline{W}'$; la seconda volta si corregge il vettore dei pesi $\underline{W}'$ utilizzando l'insieme v per ottenere il vettore dei pesi finali.

Per la scelta del numero J dei totali noti, t_X , occorre ricordare che J è il numero dei vincoli della procedura di stima e, quindi, per la sua definizione valgono le seguenti regole:

1. il valore prescelto per J non deve mai essere superiore al numero di unità campionarie presenti in ciascuno degli A domini di studio;
2. se il valore prescelto di J è troppo elevato, le differenze tra i totali noti t_X e le corrispondenti stime dirette $\tilde{\underline{t}}_{x\pi}$ sono elevate,

la procedura iterativa può non convergere ad una soluzione accettabile. L'entità di tali differenze è sostanzialmente dovuta al fatto che, in tal caso, le stime dirette sono molto variabili essendo calcolate sulla base di poche osservazioni campionarie. Quindi, nel caso in cui, per alcuni degli A domini di studio, si verifica una delle seguenti situazioni:

a) la procedura iterativa non converge, per qualsiasi valore assegnato ai parametri L U e MAXI;
b) la procedura iterativa converge, solo assegnando al parametro L valori negativi;
c) la procedura iterativa converge ad una soluzione ma, i valori calcolati delle statistiche sui coefficenti di correzione, $F\left(\underline{X}_k' \underline{\lambda}\right)$, dei pesi diretti, D_k (per $k \in s$), risultanti dalla tabella 2 del file CMS di output STIMA4 LISTING, non risultano accettabili [10];

occorre ridurre il numero, J ,dei vincoli in modo tale da non rientrare in una delle situazioni a), b) o c) appena descritte.

3.7.2. Caso delle indagini ISTAT sulle famiglie

Allo scopo di illustrare il metodo per il calcolo dei pesi base D_k è necessario descrivere brevemente il disegno di campionamento utilizzato, in genere, nelle indagini ISTAT sulle famiglie.

[10] In quanto, ad esempio, i coefficenti di correzione ottenuti ,hanno un coefficente di variazione troppo alto oppure, un valore minimo e massimo troppo distanti da 1.

Le indagini in oggetto si basano su un disegno di campionamento a due stadi con stratificazione delle unità primarie. Le unità primarie sono i comuni; le unità secondarie sono le famiglie. I comuni vengono suddivisi, nell' ambito di ciascun dominio territoriale (nella maggiorparte dei casi la regione), in strati definiti in base alla popolazione residente. In ciascun dominio territoriale gli strati presentano, approssimativamente, il medesimo ammontare di popolazione residente. Nell'ambito di ciascuno strato vengono selezionati uno o due comuni campione. La selezione avviene senza reimmissione e con probabilità proporzionale alla popolazione residente dei comuni stessi. Nell' ambito di ciascun comune campione le famiglie vengono selezionate con probabilità uguale e senza reimmisione mediante campionamento sistematico. Tutti i membri appartenenti alle famiglie selezionate sono inclusi nel campione.

Nel disegno in oggetto la probabilità di inclusione nel campione della famiglia k selezionata e rispondente nel comune i appartenente allo strato h è definita da:

$$\pi_k = \frac{n_h P_{hi}}{P_h} \frac{m_{hi}}{M_{hi}},$$

dove:

P_h indica il numero totale di persone residenti nello strato h;

P_{hi} indica il numero totale di persone residenti nel comune i dello strato h;

M_{hi} indica il numero totale di famiglie residenti nel comune i dello strato h;

m_{hi} indica il numero totale di famiglie campione selezionate nel comune i dello strato h;

n_h indica il numero di comuni selezionati nello strato h.

Tutti gli individui appartenenti alla famiglia k presentano la medesima probabilità di inclusione della famiglia. Pertanto, il peso base assegnato alla famiglia ed a ciascuno dei suoi membri è espresso da:

$$D_k = \frac{1}{\pi_k}.$$

Al fine di illustrare le modalità di costruzione dei data set SAS di input supponiamo di disporre un campione di 10.000 famiglie, tutte rispondenti all'indagine in cui i domini oggetto di studio siano le due grandi ripartizioni geografiche: Italia Settentrionale (DOMINIO =1) ed Italia Centro Meridionale (DOMINIO=2). Supponiamo, inoltre, di assumere come totali noti l'ammontare della popolazione per sesso distintamente per le due classi di età 0-20, superiore a 20.

Per soddisfare i vincoli imposti il data Set SAS NOTI.TOTALI deve pertanto avere 2 osservazioni (una per dominio) e le 5 variabili:

Nome Variabile	Contenuto della Variabile
DOMINIO	Codice di dominio, tale variabile assume valore 1 per la prima osservazione relativa all'Italia Settentrionale e valore 2 per la seconda osservazione relativa all'Italia Centro Meridionale.
TX1	Totale della popolazione di sesso maschile di età inferiore od uguale a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.
TX2	Totale della popolazione di sesso maschile di età superiore a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.
TX3	Totale della popolazione di sesso femminile di età inferiore od uguale a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.
TX4	Totale della popolazione di sesso femminile di età superiore a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.

Il data Set SAS STIME.TOTALI deve avere 2 osservazioni (una per dominio) e le 5 variabili:

Nome Variabile	Contenuto della Variabile
DOMINIO	Codice di dominio.
SX1	Stima diretta della popolazione di sesso maschile di età inferiore od uguale a 20 anni relativo alla ripartizione cui si riferisce l'osservazione. La stima in oggetto viene ottenuta sommando i pesi diretti relativi a tutti i maschi di età inferiore a 20 anni rispondenti nella ripartizione.
SX2	stima diretta del totale della popolazione di sesso maschile di età superiore a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.
SX3	Stima diretta del totale della popolazione di sesso femminile di età inferiore od uguale a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.
SX4	Stima diretta del totale della popolazione di sesso femminile di età superiore a 20 anni relativo alla ripartizione cui si riferisce l'osservazione.

Il Data set SAS DATI.CAMPIONE deve avere 10.000 osservazioni, una per ciascuna famiglia rispondente all' indagine, e le seguenti 7 variabili:

Nome Variabile	Contenuto della Variabile
CODICE	Codice (numerico) identificativo della famiglia campione, cui si riferisce l'osservazione.
DOMINIO	Codice di dominio della famiglia.
COEF	Peso diretto attribuito alla famiglia.
X1	Numero di individui della famiglia di sesso maschile di età inferiore od uguale a 20 anni.
X2	Numero di individui della famiglia di sesso maschile di età superiore a 20 anni.
X3	Numero di individui della famiglia di sesso femminile di età inferiore od uguale a 20 anni.
X4	Numero di individui della famiglia di sesso femminile di età superiore a 20 anni.

Le modalità di costruzione dei data set di input appena descritte portano a definire, mediante l'applicazione della procedura di stima, un sistema di pesi finali $\underline{W}$ che può essere utilizzato sia per ottenere stime riferite alle famiglie che stime riferite agli individui. Le stime sugli individui si ottengono associando ciascun peso finale famigliare W_k a tutti gli individui ad essa appartenenti. L'utilizzazione di un unico sistema di pesi per ottenere stime riferite sia alle famiglie che agli individui garantisce il fatto che le stime riferite alle famiglie non siano in contraddizione con quelle riferite agli individui. Inoltre il sistema dei pesi finali applicato per ottenere stime riferite agli individui garantisce il rispetto della condizione di uguaglianza tra i totali noti, al livello delle due ripartizioni geografiche, della popolazione residente per sesso e classi di età e le corrispondenti stime campionarie.

3.7.3. Caso delle indagini ISTAT sulle imprese

Allo scopo di illustrare il metodo per il calcolo dei pesi base D_k è necessario descrivere brevemente il disegno di campionamento utilizzato, in genere, nelle indagini ISTAT sulle imprese.

Le indagini in oggetto si basano su un disegno di campionamento ad uno stadio stratificato. Le imprese vengono suddivise, nell' ambito di ciascun dominio territoriale (nella maggiorparte dei casi la regione), in strati definiti in base alla classe di attività economica ed alla classe di addetti di ciascuna impresa. Nell' ambito di ciascuno strato le imprese campione vengono selezionate senza reimmissione e con probabilità uguali.

Nel disegno in oggetto la probabilità di inclusione nel campione dell'impresa k selezionata e rispondente ed appartenente allo strato h è definita da:

$$\pi_k = \frac{m_h}{M_h},$$

dove:

M_h indica il numero totale di imprese appartenenti allo strato h;

m_h indica il numero totale di imprese campione selezionate nello strato h.

Pertanto, il peso base assegnato a tutte le imprese rispondenti dello strato h è espresso da:

$$D_k = \frac{1}{\pi_k}.$$

Al fine di illustrare le modalità di costruzione dei data set SAS di input supponiamo di disporre un campione di 1.000 imprese, tutte rispondenti all'indagine in cui i domini oggetto di studio siano le due grandi ripartizioni geografiche: Italia Settentrionale (DOMINIO =1) ed Italia Centro Meridionale (DOMINIO=2). Supponiamo, inoltre, di non disporre di informazioni note relative al periodo di riferimento dell' indagine sulla popolazione delle imprese. Su tale popolazione conosciamo invece le informazioni desunte dall' archivio da cui le 1.000 imprese sono state selezionate. Supponiamo, quindi di volere assumere come totali noti, per ciascun dominio geografico di riferimento, il numero delle imprese ed il numero degli addetti ad esse appartenenti per 2 classi di attività economica: la prima relativa alle imprese di produzione, la seconda relativa ai servizi; le quantità in oggetto sono naturalmente riferite alla data di formazione dell' archivio.

Per soddisfare i vincoli imposti il data Set SAS NOTI.TOTALI deve pertanto avere 2 osservazioni (una per dominio) e le 5 variabili:

Nome Variabile	Contenuto della Variabile
DOMINIO	Codice di dominio, tale variabile assume valore 1 per la prima osservazione relativa all'Italia Settentrionale e valore 2 per la seconda osservazione relativa all'Italia Centro Meridionale.
TX1	Totale imprese di produzione presenti nell' archivio relative al dominio a cui si riferisce l'osservazione.
TX2	Totale addetti (desunto dall' archivio) appartenenti alle imprese di produzione presenti nell' archivio relative al dominio a cui si riferisce l'osservazione.
TX3	Totale imprese di servizio presenti nell' archivio relative al dominio a cui si riferisce l'osservazione.
TX4	Totale addetti (desunto dall' archivio) appartenenti alle imprese di servizi presenti nell' archivio relative al dominio a cui si riferisce l'osservazione.

Il data Set SAS STIME.TOTALI deve avere 2 osservazioni (una per dominio) e le 5 variabili:

Nome Variabile	Contenuto della Variabile
DOMINIO	Codice di dominio, tale variabile assume valore 1 per la prima osservazione relativa all'Italia Settentrionale e valore 2 per la seconda osservazione relativa all'Italia Centro Meridionale.
SX1	Stima diretta del totale imprese di produzione presenti nell' archivio relative al dominio a cui si riferisce l'osservazione. La quantità in oggetto si ottiene sommando i pesi diretti di tutte le imprese di produzione rispondenti all'indagine ed appartenenti al dominio di riferimento.
SX2	Stima diretta del totale addetti (desunto dall' archivio) appartenenti alle imprese di produzione presenti nell' archivio relative al dominio a cui si riferisce l'osservazione. La quantità in oggetto si ottiene moltiplicando, per ciascuna delle aziende di produzione rispondenti il peso diretto per il numero degli addetti della medesima azienda riportato sull'archivio di selezione del campione. Successivamente si sommano i prodotti così ottenuti su tutte le imprese di produzione rispondenti all'indagine ed appartenenti al dominio di riferimento.
SX3	Stima diretta del totale imprese di servizio presenti nell' archivio relative al dominio a cui si riferisce l'osservazione.
SX4	Stima diretta del totale addetti (desunto dall' archivio) appartenenti alle imprese di servizi presenti nell' archivio relative al dominio a cui si riferisce l'osservazione.

Il Data set SAS DATI.CAMPIONE deve avere 1.000 osservazioni, una per ciascuna impresa rispondente all' indagine, e le seguenti 7 variabili:

Nome Variabile	Contenuto della Variabile
CODICE	Codice (numerico) identificativo della imprese campione, cui si riferisce l'osservazione.
DOMINIO	Codice di dominio dell' impresa.
COEF	Peso diretto attribuito all' impresa.
X1	Variabile indicatrice che assume valore 1 o 0. X1=1 se l'impresa, nell' archivio utilizzato per la selezione del campione, risulta essere classificata come impresa di produzione; X1=0, altrimenti.
X2	X2 = al numero di addetti dell'impresa (desunto dall' archivio) se l'impresa , nell' archivio utilizzato per la selezione del campione, risulta essere classificata come impresa di produzione; X2=0, altrimenti.
X3	Variabile indicatrice che assume valore 1 o 0. X3=1 se l'impresa, nell' archivio utilizzato per la selezione del campione, risulta essere classificata come impresa di servizi; X3=0, altrimenti.
X4	X4 = al numero di addetti dell'impresa (desunto dall' archivio) se l'impresa , nell' archivio utilizzato per la selezione del campione, risulta essere classificata come impresa di servizi; X4=0, altrimenti.

Da quanto appena descritto, risulta chiaro che: il valore da assegnare alle variabili X1, X2, X3 ed X4 per ciascuna delle imprese rispondenti all'indagine viene determinato sulla base della classe di attività economica e del numero di addetti che sono riportati, per ciascuna delle imprese rispondenti all'indagine, sull'archivio di estrazione del campione.

Note

Il lavoro è frutto della collaborazione degli Autori. Per quanto riguarda la sua stesura, S. Falorsi ha redatto i paragrafi 2.2, 2.3, 3.1, 3.2, 3.4, 3.5 e 3.7.3. I rimanenti paragrafi sono stati redatti da P.D. Falorsi.

RIFERIMENTI BIBLIOGRAFICI

ALEXANDER C. H. (1990) "Incorporating Person Estimates into Household Weighting Using Various Models for Coverage", Annual Research Conference , USA, 1990, U.S. Department of Commerce, Bureau of the Census, pp. 445-461.

BINDER D. A., THEBERGE A. (1988) " Estimating the Variance of Raking-Ratio Estimators", The Canadian Journal of Statistics, vol. 16, pp. 47-55.

DARROCH J. N., RATCLIFF D. (1972) "Generalized Iterative Scaling for Log-Linear Models", The Annals of Mathematical Statistics, vol. 43, pp. 1470-1480.

DEVILLE J. C., SARNDAL C. E. (1992) " Calibration Estimators in Survey Sampling", Journal of the American Statistical Association , vol. 87, pp. 376-382.

FALORSI P. D., FALORSI S., RUSSO A. (1991) "Indagine Statistica sulle Condizioni di Salute della Popolazione e sul Ricorso ai Servizi Sanitari", Note e Relazioni n.21., ISTAT 1991.

FALORSI P. D., FALORSI S., RUSSO A. (1993) "Indagine Multiscopo sulle Famiglie - Anni 1987 - 1991: Obiettivi, Disegno e Metodologia dell'Indagine", vol. 1 , ISTAT 1993.

HORVITZ D. G., THOMPSON D. J. (1952) "A generalization of sampling without replacement from a finite universe", Journal of the American Statistical Association , vol. 47, pp. 663-685.

ISTAT, (1989), "MANUALE DI TECNICHE D'INDAGINE 4- tecniche di campionamento teoria e pratica", anno 1989, Note e relazioni n.1.

ISAKI C. T., FULLER W. A. (1982) " Survey Design Under the Regression Superpopulation Model" , Journal of the American Statistical Association , vol. 77, pp. 89-96.

SARNDAL C. E., SWENSSON B., WRETMAN J. (1992) *Model Assisted Survey Sampling*, Springer - Verlag, New York.