MAUSS (MULTIVARIATE ALLOCATION OF UNITS IN SAMPLING SURVEYS): UN SOFTWARE GENERALIZZATO PER RISOLVERE IL PROBLEMA DELL'ALLOCAZIONE CAMPIONARIA NELLE INDAGINI ISTAT

di

ROBERTO DI GIUSEPPE, PATRIZIA GIAQUINTO, DANIELA PAGLIUCA [(*)]

[(*)] Patrizia Giaquinto e Daniela Pagliuca hanno redatto le Parti I e II (Parte 1: Patrizia Giaquinto ha redatto i paragrafi 1.1 e 3. Daniela Pagliuca ha redatto i paragrafi rimanenti; La Parte II è stata redatta in comune). Roberto Di Giuseppe ha redatto la documentazione progettuale (Parte III).

Premessa

Il presente documento ha la finalità di divulgare i risultati attualmente ottenuti nell'ambito del progetto Istat per lo sviluppo di un software generalizzato, denominato MAUSS[1] (acronimo per *Multivariate Allocation of Units in Sampling Surveys*).

Il progetto si pone come obiettivo lo sviluppo di uno strumento software generalizzato per la determinazione dell'allocazione campionaria nel caso multivariato e per più domini di stima, applicabile alle indagini con disegni ad uno e a due stadi di campionamento. Il progetto nasce dalla esigenza di potenziare un prototipo sviluppato in precedenza in Istat allo scopo di determinare l'allocazione delle unità in campioni ad un unico stadio di campionamento, i cui principali referenti, sia in quanto autori del prototipo, sia per la metodologia implementata, sono Marco Ballin e Piero Demetrio Falorsi.

Tale progetto di sviluppo è svolto nell'ambito della direzione DCMT e costituisce una attività programmata dall'unità MTS/F, che si occupa dello sviluppo di software generalizzati per la produzione statistica, la cui responsabile è Daniela Pagliuca. L'unità afferisce al servizio MTS "Servizio metodologie, tecnologie e software per la produzione statistica", il cui responsabile è Giulio Barcaroli. L'attività è stata svolta in collaborazione con il servizio PSM "Progettazione e supporto metodologico nei processi di produzione statistica", di cui è responsabile Piero Demetrio Falorsi.

Il progetto è articolato in due fasi successive:

1. La prima fase, prossima alla conclusione, riguarda la realizzazione del software generalizzato MAUSS, per la determinazione dell'allocazione campionaria multivariata per disegni ad unico stadio di campionamento. MAUSS amplia le potenzialità previste dal prototipo iniziale: il prototipo infatti viene tuttora utilizzato in Istat, ma comprende opzioni non attivate e produce alcuni risultati non convalidati. Esso pertanto è stato rivisto, corretto ed esteso nelle funzionalità. Il software generalizzato è dotato di una interfaccia *user-friendly* e di opportuni controlli logici, e assicura una buona flessibilità, comprendendo alcune opzioni aggiuntive relative alla valutazione di possibili alternative. Allo stato attuale è divulgata la versione beta del software.

2. La seconda fase è programmata per rispondere alle esigenze degli utenti che lavorano con disegni campionari a due stadi, utilizzati principalmente nelle indagini Istat che si occupano delle famiglie. In tal senso è prevista la realizzazione di una nuova funzione di MAUSS e attualmente è stato sviluppato un primo prototipo software, che implementa un modulo alla base di questa nuova funzione.

Il documento ha una struttura che consta di tre parti distinte.

La prima parte (comprendente i paragrafi 1, 2 e 3) è descrittiva, in quanto concerne la fase progettuale descritta al punto 1, che si trova ad uno stadio avanzato: essendo già stabilita la definizione del software MAUSS in termini di funzioni-utente, il software implementato viene illustrato in modo semplice e comprensibile.

[1] Paolo Floris ha sviluppato il software fino ad Ottobre 2003. Roberto Di Giuseppe ha successivamente ottenuto tale incarico ed è attualmente il responsabile e referente per il software generalizzato MAUSS; ha prodotto la documentazione progettuale ed ha modificato il software, basandosi sulle verifiche eseguite da Patrizia Giaquinto, che a ha la responsabilità della fase di test del software.

La Parte II (paragrafi 4, 5 e 6) è invece più specialistica e l'esposizione risulta necessariamente di carattere tecnico, in quanto si riferisce alla seconda fase progettuale, che, come si è detto, riguarda la progettazione e implementazione di una nuova funzione in MAUSS.
Allo stato attuale è in corso l' analisi dei requisiti e la progettazione preliminare che ha prodotto un prototipo iniziale. Il lettore interessato alla problematica generale può quindi consultare solo i paragrafi 4 e 5, tralasciando il paragrafo 6 relativo alle specifiche della progettazione; per facilitare la presentazione infatti sono stati inseriti nel documento alcuni passi di programmazione implementati tramite il linguaggio SAS.

La terza e ultima parte consiste negli allegati progettuali del software MAUSS: essa pertanto è rivolta a quanti siano eventualmente interessati alla documentazione progettuale vera e propria.

Parte I

La Parte I descrive il software generalizzato MAUSS, concepito per determinare l'allocazione campionaria per disegni ad unico stadio di campionamento.

Nel paragrafo 1 sono riportati alcuni elementi introduttivi al problema e utili per coloro che volessero approfondire gli aspetti generali e metodologici alla base del software; nel paragrafo 2 e 3 sono invece rispettivamente analizzati il software MAUSS (paragrafo 2), e le stampe che esso produce (paragrafo 3).

Figura 1: La schermata principale del software MAUSS

1. Il problema generale e la soluzione metodologica

Le indagini campionarie condotte in Istat generalmente prendono in considerazione molte variabili di interesse e frequentemente devono produrre stime di parametri riferiti ad un numero elevato di domini di stima. Nella determinazione dell'allocazione campionaria, è perciò necessario individuare la soluzione ottima, in presenza di diverse finalità da considerare congiuntamente.

Nel caso in cui siano disponibili informazioni sulla variabilità negli strati, esse possono essere utilizzate nella ricerca di una soluzione che assicuri prefissati livelli di precisione per le stime di

interesse; se si fa riferimento a più domini di stima, la dimensione campionaria va determinata considerando che le stime devono essere prodotte a livello dei diversi domini di stima e che l'accuratezza delle stime deve essere garantita simultaneamente per tali sottopopolazioni.

In Istat è stato implementato – già nel 1998 - un software (Falorsi P.D., Ballin, De Vitiis, Scepi, 1998) sviluppato a livello prototipale, in cui si è adottata la soluzione proposta da Bethel nel 1989 (Bethel, 1989) per il calcolo dell'allocazione ottima nel caso multivariato, generalizzandola al caso analogo ma per più domini di stima.

Per chiarimenti sui concetti relativi a tale problematica è consigliabile la lettura di (Falorsi P.D., Ballin, De Vitiis, Scepi, 1998). In questa sede è opportuno puntualizzare le definizioni più rilevanti .

Il software determina l'allocazione campionaria sulla base di informazioni definite a livello di *strato*, dove con "strato" si intende una parte della popolazione all'interno della quale le unità sono in qualche modo omogenee. E' utile evidenziare che lo *strato* può corrispondere ad una variabile di stratificazione scelta dal responsabile di indagine o ad un incrocio di modalità di variabili che definiscono la stratificazione (ad esempio nelle indagini sulle imprese lo strato potrebbe essere definito dall'incrocio delle modalità di una variabile territoriale con le modalità di una variabile di classificazione economica).

Le stime dei parametri di interesse difficilmente sono riportate solo a livello dell'intera popolazione: le indagini solitamente forniscono stime anche per diverse sottopopolazioni, dette *domini di stima*. I *domini di stima* possono ottenersi come unione di *strati* e, in tal modo, gli strati rappresentano la minima partizione atta a determinare il livello di interesse al quale riportare le stime finali (provincia, regione,..).

Il metodo originale (Bethel, 1989) è finalizzato alla determinazione della dimensione campionaria ottimale che assicuri l'ottenimento delle stime dei parametri di interesse con il livello di precisione desiderato, in un'ottica multivariata e considerando disegni ad un unico stadio stratificato e un solo dominio di stima.

Tale metodologia è stata implementata in modo esteso invece per consentire la trattazione del caso di diversi domini di stima contemporaneamente, determinando la dimensione campionaria *minima* richiesta (tale soluzione è tuttavia meno efficiente di quella ottenibile considerando i diversi domini di stima separatamente).

La soluzione ottima è individuata in modo iterativo, mediante l'implementazione dell'algoritmo di Chromy (Chromy, 1987).

Il prototipo è dunque specifico delle sole indagini ad uno stadio, ma, come già specificato, si è delineata la necessità di contemplare anche il caso dei due stadi per supportare in particolare le indagini sulle famiglie in Istat.

Si riporta di seguito un prospetto riassuntivo delle strategie campionarie che si utilizzano tipicamente nelle indagini Istat:

1. *ad uno stadio stratificato* – adottato da molte indagini sulle imprese;
2. *a due stadi* per le indagini sulle imprese (adottato raramente);
3. *a due stadi* per le indagini sulle famiglie, *con stratificazione delle unità di primo stadio,* generalmente i comuni, e le famiglie come unità di secondo stadio – adottato da molte indagini sulle famiglie;
4. *ad uno stadio stratificato* per le indagini sulle famiglie (adottato meno frequentemente).

Il primo disegno è ampiamente soddisfatto dall'implementazione del software generalizzato MAUSS che attualmente è distribuito in versione beta; la seconda fase progettuale è alla base del disegno 3.

1.1 Cenni metodologici

Il problema dell'allocazione del campione negli strati consiste nella determinazione della ampiezza campionaria minima, all'interno dei singoli domini di stima, sotto il vincolo di contenimento della variabilità degli stimatori dei parametri di interesse entro livelli prefissati.

La metodologia impiegata a tale scopo costituisce in pratica una estensione del metodo di Neyman al caso di più variabili, e adotta poi come metodo di risoluzione una generalizzazione della proposta di Bethel (1989).

Essa viene qui riportata in maniera molto schematica ed essenziale, seguendo l'impostazione utilizzata in (Falorsi P.D., Ballin, De Vitiis, Scepi, 1998).

Si indichi con

$$V'_{p,k_d} = \sum_{h=1}^{H_{k_d}} \frac{N_h^2}{n_h} S_{p,h}^2 - \sum_{h=1}^{H_{k_d}} N_h S_{p,h}^2 = V_{p,k_d} + V_{0p,k_d} \qquad (p=1,\ldots, P;\ d=1,\ldots, D;\ k_d = 1,\ldots K_d)$$

la varianza della stima del totale della variabile p per il generico dominio k_d del tipo di dominio d. In particolare, V_{0p,k_d} denota la parte indipendente dall'allocazione.

Analogamente, si indichi con

$$V^*_{p,k_d} \qquad (p=1,\ldots, P;\ d=1,\ldots, D;\ k_d = 1,\ldots K_d)$$

il suo estremo superiore.

Si definisca inoltre una funzione di costo:

$$C' = C_0 + C = C_0 + \sum_{h=1}^{H_{k_d}} C_h n_h , \qquad (h = 1,...., H_{kd})$$

in cui C_0 rappresenta una quota di costo fisso, indipendente dall'allocazione, mentre C è il costo variabile in relazione alla dimensione campionaria. H_{kd} è infatti il numero di strati nel dominio k_d e n_h è la rispettiva numerosità del campione. A quest'ultima viene associato un costo C_h relativo ad ogni singolo strato.

Nella metodologia in questione, il problema dell'allocazione si può formalizzare come segue:

$$\min C \quad \text{sotto i vincoli } V'_{p,k_d} \le V^*_{p,k_d} \qquad (p=1,\ldots, P;\ d=1,\ldots, D;\ k_d = 1,\ldots K_d). \qquad (1)$$

La proposta di Bethel parte dalla considerazione che la (1) può essere vista come un problema di minimo vincolato. Allo scopo, occorre porre:

$$l_{p,k_d,h} = \frac{N_h^2 S_{p,h}^2 \rho_{k_d,h}}{\sum_{h=1}^{H_{k_d}} N_h S_{p,h}^2 \rho_{k_d,h} + V^*_{p,k_d}} \qquad (2)$$

e

$$\rho_{k_d,h} = \begin{cases} 1 \text{ se } h \in k_d \\ 0 \text{ altrimenti} \end{cases}$$

e definire un indice r (r = 1,…, R), i cui valori stabiliscono una corrispondenza biunivoca con i valori ordinati del vettore contenente i tre indici d, k_d, p, in modo tale che $R = P\sum_{d=1}^{D} K_d$.

Il sistema dei vincoli (1) può essere allora riformulato come:

$$\sum_{h=1}^{H_{k_d}} l_{p,k_d,h} x_h \le 1 \qquad (p=1,\ldots, P;\ d=1,\ldots, D;\ k_d = 1,\ldots K_d) \qquad (3).$$

Ponendo, inoltre, $x_h = \begin{cases} 1/n_h \text{ se } n_h \ge 1 \\ \infty \text{ altrimenti} \end{cases}$,

la funzione obiettivo da minimizzare diventa:

$$f(x) = \sum_{h=1}^{H_{k_d}} C_h \Big/ x_h \quad \text{dove } x = (x_1,\ldots,x_{Hkd})' \qquad (4).$$

La soluzione ottima per la (4) risulta essere:

$$x_h^* = \sqrt{C_h} \Big/ \left(\sqrt{\sum_{r=1}^{R} \mu_r^* l_{r,h}} \sum_{k=1}^{H} \sqrt{C_k \sum_{r=1}^{R} \mu_r^* l_{r,k}} \right), \qquad (5)$$

in cui $\mu_r^* = \lambda_r^* \Big/ \sum_{r=1}^{R} \lambda_r^*$ e $\sum_{r=1}^{R} \mu_r^* = 1$.

L'algoritmo proposto da Bethel per il calcolo della soluzione ottima x_h^* è di tipo iterativo, per cui ad ogni passo v (posto inizialmente pari a 1), la numerosità campionaria viene aumentata incrementando la funzione obiettivo $f(\mathbf{x}^{(v)}) \ge f(\mathbf{x}^{(v-1)})$ fino al soddisfacimento di tutti i vincoli. Bethel dimostra anche che l'algoritmo converge, ma, data la complessità dal punto di vista computazionale della sua proposta, si è scelto di implementare nel prototipo software, e analogamente in MAUSS, l'alternativa presentata da Chromy nel 1987.

L'algoritmo di Chromy, oltre ad essere più immediato del precedente in termini di implementazione, converge anche più velocemente verso la soluzione ottima.

Esso parte dalla definizione della matrice $L = \{l_{r,h}\}$ di dimensione R x H, con elementi dati dalla (2).

Essendo anche questo un algoritmo iterativo, al primo passo si calcola il valore di **x** in base alla (5), ponendo ogni elemento di μ pari a 1/R. Se la soluzione soddisfa tutti i vincoli l'algoritmo si arresta, altrimenti si calcola $\mathbf{x}^{(\nu)}$ in corrispondenza del valore $\boldsymbol{\mu}^{(\nu)}$ ottenuto come di seguito:

$$\mu_r^{(\nu)} = \mu_r^{(\nu-1)} (\mathbf{l}_r \mathbf{x}(\mu^{(\nu-1)}))^2 \Big/ \sum_{r=1}^{R} \mu_r^{(\nu-1)} (\mathbf{l}_r \mathbf{x}(\mu^{(\nu-1)}))^2 \qquad 1 \le r \le R.$$

2. Il software generalizzato per le indagini ad uno stadio stratificato e le sue funzioni

Questo paragrafo illustra la versione beta del software generalizzato MAUSS, attualmente implementata e distribuita tramite le pagine internet messe a disposizione dall'unità MTS/F per permettere di effettuare il download dei software generalizzati prodotti[2] .

Il software MAUSS è stato sviluppato utilizzando il SAS SYSTEM v8 per Microsoft Windows, ovvero un package di uso generale che incorpora statistiche e procedure di analisi dei dati. Per utilizzarlo è pertanto necessario che sia installato il sistema SAS versione 8 ed in particolare i moduli: SAS Language and Macro-facility, SAS IML Language, SAS STAT.

Lo spazio sul disco fisso richiesto per l'installazione è di circa 4 MB ed è consigliabile una memoria di almeno 64 MB. Il tempo d'esecuzione della procedura è legato, ovviamente, alla velocità del processore installato e alla dimensione e complessità dei dati da elaborare.

All'utente è richiesto di costruire preventivamente i due data-set di input - il data-set degli ***strati*** e il data-set dei ***vincoli*** – e dopo l'elaborazione produce in output l'allocazione campionaria, ovviamente definita sulla base delle informazioni di input, e produce alcune stampe.

Nei prossimi due paragrafi sono illustrati i due data-set di input richiesti da MAUSS (le informazioni contenute sono le stesse necessarie per il funzionamento del prototipo software sviluppato in precedenza) e vengono descritte le funzioni implementate: in particolare nel paragrafo 2.1 si descrivono i due data-set SAS e nel paragrafo 2.2 le funzioni del software. Le stampe vengono descritte nel successivo paragrafo 3.

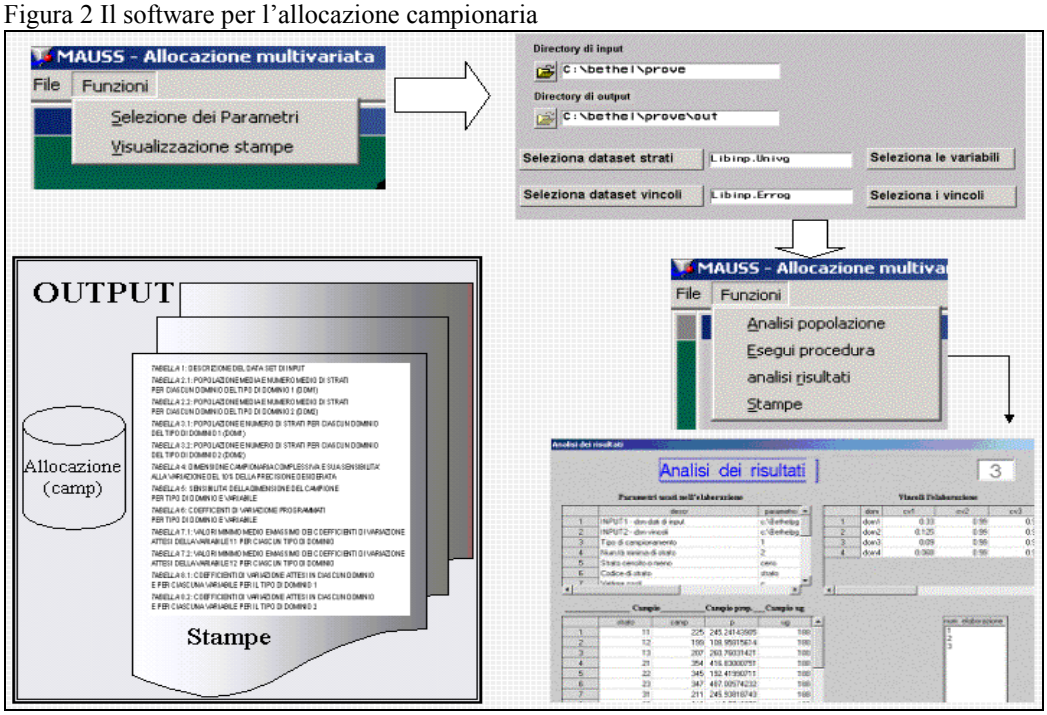
Come premessa, in figura 2 viene schematizzata la struttura implementata.

Il software è definito da **due funzioni principali**:

a. La selezione dei parametri
b. La visualizzazione delle stampe

[2] Il software è disponibile: Via intranet (per utenti istat): http://intranet/ (selezionare: "Prodotti e Applicazioni on-line. Software Generalizzati" e da qui selezionare "**MTS-F**: Software Generalizzati per la Produzione Statistica (Area Download e Informazioni)"). Via internet (per utenti esterni all'istat): http: //www.istat.it/Metodologi/index.htm (selezionare "Metodi e Software per indagini statistiche").

Figura 2 Il software per l'allocazione campionaria



a. La selezione dei parametri

MAUSS è dotato di una adeguata interfaccia grafica che facilita l'utente nella selezione dei dati di input necessari ad attivare le elaborazioni delle informazioni di interesse. Le elaborazioni previste dalla voce "Funzioni" (figura 2) permettono di memorizzare diverse alternative variando i vincoli, e visualizzare le allocazioni che si ottengono (si veda paragrafo 2.2).

L'allocazione ottenuta da ogni singola elaborazione è poi confrontabile con una allocazione proporzionale o uguale negli strati ("Analisi dei Risultati" in figura 2, si veda paragrafo 2.2).

b. La visualizzazione delle stampe

Il software produce sostanzialmente due tipi di stampe:

- il primo costituisce un prospetto informativo generale della popolazione oggetto di studio
- il secondo, oltre alle informazioni di cui sopra, riporta i risultati ottenuti applicando la procedura di allocazione del campione in strati (si veda paragrafo 3).

2.1 I data-set di input

In figura 3 viene mostrato un esempio di data-set degli strati contenente le informazioni fondamentali per determinare l'allocazione tramite MAUSS.

Figura 3: un esempio di data-set di input - il data-set degli strati

Column Name	Type	Length	Format	Informat	La
STRATO	Text	8			
N	Number	8			
S1	Number	8			
S2	Number	8			
S3	Number	8			
S4	Number	8			
S5	Number	8			
M1	Number	8			
M2	Number	8			
M3	Number	8			
M4	Number	8			
M5	Number	8			
DOM1	Text	8			
DOM2	Text	8			
DOM3	Text	8			
DOM4	Text	8			
C	Number	8			
cens	Number	8			

Il data-set quindi deve contenere un record per ciascuno strato con le seguenti informazioni:

- STRATO (codice di strato)
- N (numero di unità dello strato della popolazione)
- M1, M2,....,MP (medie delle variabili d'interesse)
- S1, S2,....,SP (s.q.m. delle variabili d'interesse)
- DOM1,DOM2,..,DOMD (codici di tipo di dominio di stima)
- C (costo unitario di rilevazione per strato - default C=1)
- CENS (indica se lo strato deve essere censito o meno; cens=1 se censito; cens=0 se camp.- default cens=0)

Può inoltre contenere una variabile GRUPPO che indica una suddivisone della intera popolazione in sottopopolazioni

Figura 4: un esempio di data-set di input - il data-set dei vincoli

Column Name	Type	Length	Format	Informat	Label
DOM	Text	8			
CV1	Number	8			
CV2	Number	8			
CV3	Number	8			
CV4	Number	8			
CV5	Number	8			

Il secondo data-set (figura 4) deve contenere dei record con le seguenti informazioni:

- DOM (codice tipologia di dominio)
- CV1, CV2,....,CVP (coefficienti di variazione desiderati per la stima di M1,M2,...,MP)

Tra i controlli attivati in MAUSS, si evidenzia la verifica della corrispondenza fra il numero di coefficienti di variazione, nel data-set dei vincoli, e il numero di medie e degli s.q.m, nel data-set degli strati.

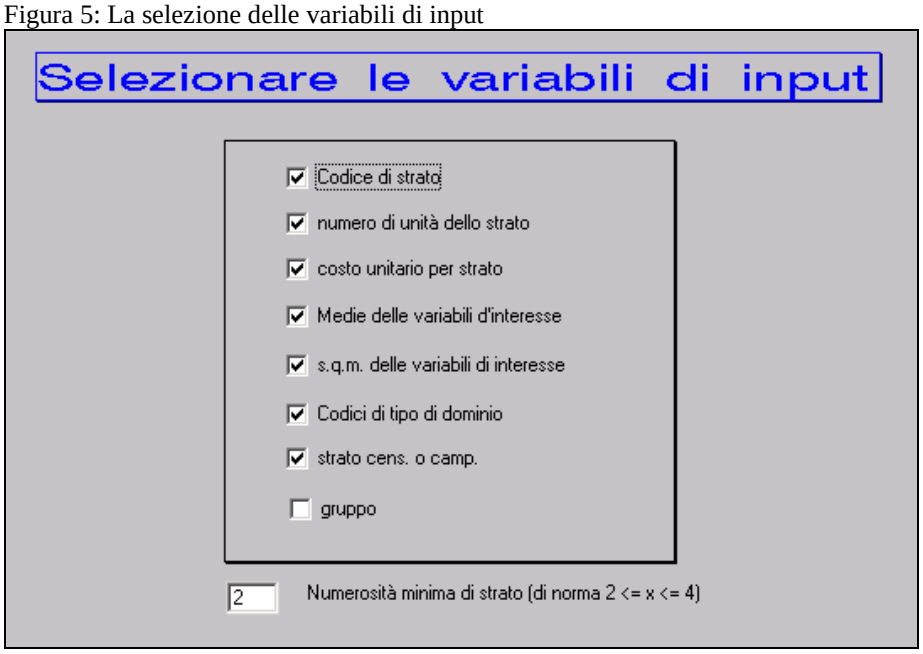
2.2 La selezione dei parametri

La funzione che permette di selezionare i parametri di input è quella che attiva l'elaborazione della procedura di allocazione campionaria. In mancanza di tale requisito (scelta dei data-set iniziali) le esecuzioni rimangono disattivate.

Nel seguito, per descrivere le funzionalità previste dal software a titolo esemplificativo, si mostrano alcune maschere interattive, in modo che si abbia una idea concreta del software e si visualizzi il livello di implementazione raggiunto.

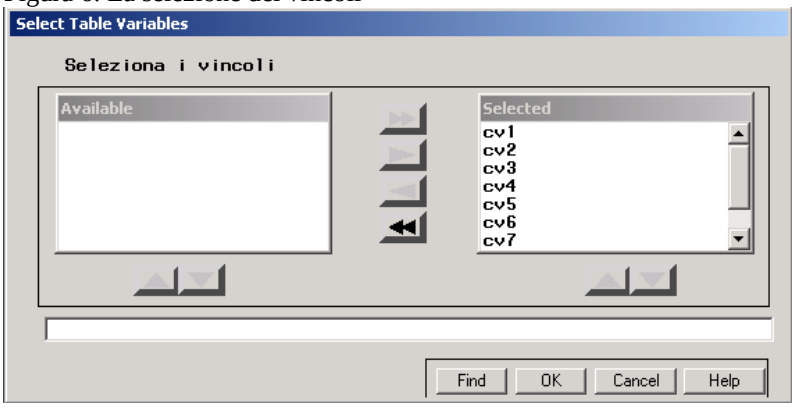
Per selezionare le variabili del data-set degli strati, occorre scegliere la cartella e il data-set tramite apposita maschera, e successivamente, si devono associare le variabili del data-set alle variabili logiche richieste dal software, così come mostrato in Figura 5:

Figura 5: La selezione delle variabili di input



Analogamente per il data-set dei vincoli, dopo avere selezionato la cartella e il data-set si devono individuare i vincoli (CV1, …..,CVP). In Figura 6 viene mostrata l'interfaccia che appare dopo aver scelto le variabili relative ai vincoli.

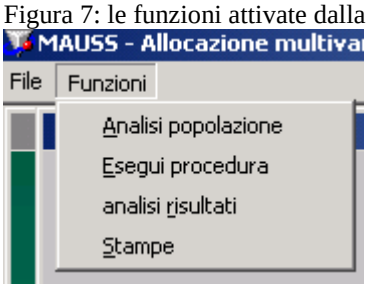
Figura 6: La selezione dei vincoli



Successivamente alle selezioni, vengono attivate le seguenti possibilità:

1. effettuare l'analisi della popolazione
2. eseguire la procedura
3. analizzare i risultati di campionamento (condizionata all'esecuzione della procedura)
4. procedere con le stampe

Figura 7: le funzioni attivate dalla selezione dei parametri

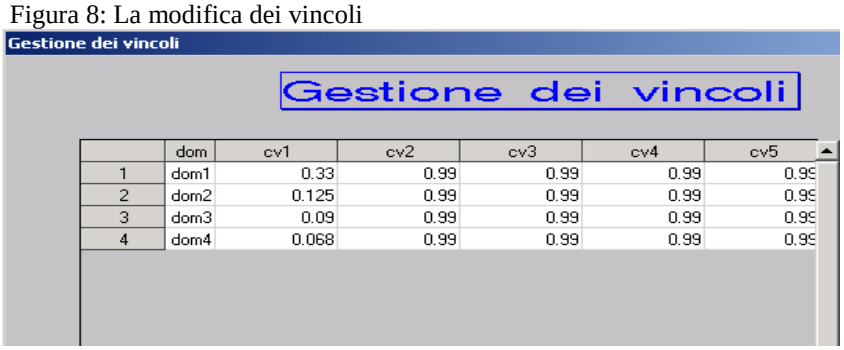


E' importante sottolineare il fatto che ogni elaborazione effettuata con lo stesso data-set degli strati e con vincoli diversi viene memorizzata dal software.

Infatti è possibile variare il valore numerico dei vincoli (CV1, ….., CVP) tramite una funzione "***modifica dei vincoli***", in modo che - fermo restando il data-set degli strati - sia possibile ripetere l'elaborazione e ottenere un nuovo risultato sulla base dei nuovi vincoli.

Per effettuare la modifica si utilizza una interfaccia quale quella riportata in figura 8.

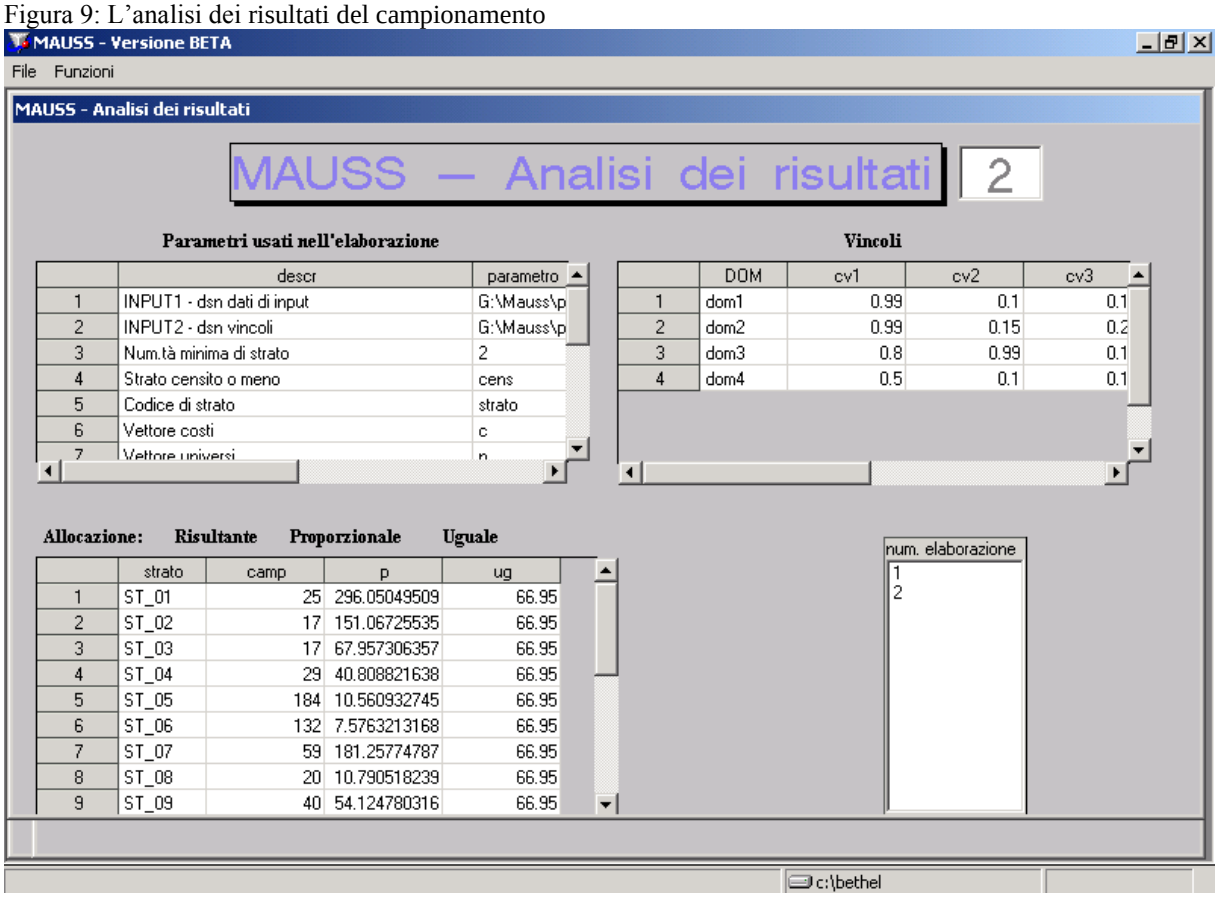
Figura 8: La modifica dei vincoli



	dom	cv1	cv2	cv3	cv4	cv5
1	dom1	0.33	0.99	0.99	0.99	0.99
2	dom2	0.125	0.99	0.99	0.99	0.99
3	dom3	0.09	0.99	0.99	0.99	0.99
4	dom4	0.068	0.99	0.99	0.99	0.99

Per ogni elaborazione sullo stesso data-set degli strati vengono dunque registrate tutte le informazioni utili, quali i vincoli utilizzati, i risultati dell'allocazione conseguita e le stampe ottenute. Tali informazioni sono anche rese disponibili e visualizzabili tramite la funzione "Analisi dei risultati" che, come mostrato in figura 7, è una di quelle attivate dalla funzione selezione dei parametri. Ovviamente l"Analisi dei risultati" è abilitata solo dopo l'esecuzione vera e propria ("Esegui procedura"**)**.

Figura 9: L'analisi dei risultati del campionamento



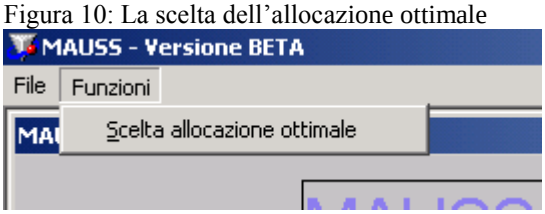
Come appare in figura 9, la maschera visualizza quattro sezioni differenti. La prima sezione, in alto a sinistra, riassume i parametri utilizzati dalla procedura, mentre la seconda, sulla destra, è relativa ai vincoli selezionati.

Nella sezione riportata in basso, sempre sulla destra, sono presenti degli indicatori numerici, che corrispondono alle elaborazioni effettuate sullo stesso data-set degli strati ("***num. elaborazione***"). Posizionandosi su un indicatore specifico, si richiama l'elaborazione di riferimento, condizionata ai vincoli riportati, per la quale è possibile visualizzare l'allocazione corrispondente. In figura 9 è possibile, ad esempio, visualizzare le informazioni relative alla seconda elaborazione.

In particolare, nella sezione sulla sinistra in basso, per ogni singolo strato viene riportato il risultato della procedura, ovvero l'allocazione ottenuta con il metodo di Bethel ("***camp"***), l'allocazione *proporzionale* ("***p***") e quella *uguale* negli strati ("***ug"***). E' quindi possibile istituire per ciascuna elaborazione un confronto immediato fra i tre risultati.

Questa informazioni facilitano all'utente la scelta di quella che ritiene essere l'allocazione migliore ottenuta tra tutte le elaborazioni compiute. Tramite la funzione "***scelta allocazione ottimale***"è pertanto possibile cancellare i risultati e le stampe di tutte le altre elaborazioni. Tale funzione è attivata dalla voce "Funzioni" nella schermata "Analisi dei risultati" (figura 10):

Figura 10: La scelta dell'allocazione ottimale



3. Le stampe

Nel software MAUSS sono previste sostanzialmente due categorie di stampe, prodotte dalle due funzioni "Analisi della popolazione" e "Esegui procedura" (si veda figura 7).

Analisi della popolazione

L'analisi della popolazione produce tre stampe che schematizzano le caratteristiche della popolazione oggetto di studio.

La prima stampa - "***Informazioni generali***" - contiene in un'unica tabella informazioni essenziali quali la numerosità della popolazione, il numero delle variabili rilevate, il numero di strati nei quali la suddetta popolazione viene complessivamente ripartita, il numero di differenti tipi di dominio considerati.

A seconda che gli strati siano stati censiti o campionati (variabile CENS nel data-set degli strati, si veda paragrafo 2.1), vengono riportate altre informazioni quali la numerosità della popolazione da censire o da campionare ed il numero degli strati che si è scelto di censire o di campionare.

La seconda e la terza stampa - "***Popolazione media e numero medio di strati per ciascun dominio del tipo di dominio***" e "***Popolazione e numero di strati per ciascun dominio del tipo di dominio***" – riportano tali informazioni per ciascun tipo di dominio.

Esegui procedura

Nel momento in cui viene eseguita la procedura di allocazione, i risultati vengono stampati a video in automatico in un prospetto suddiviso in tre sezioni informative.

- La prima sezione replica le tabelle dell'"Analisi della popolazione", poiché l'utente potrebbe decidere di eseguire la procedura senza aver stampato in precedenza le informazioni riassuntive della popolazione; in tal modo si fornisce comunque la possibilità di richiamarle direttamente con l'"Esegui procedura" per ottimizzare l'interpretazione dei risultati ottenuti.

- La seconda sezione è relativa allo studio della **sensibilità del campione** e consta di due tabelle. Per "*sensibilità campionaria*" si intende la variazione da apportare alla dimensione campionaria risultante, qualora si verifichi una variazione nella precisione della stima. Ad un aumento della precisione, infatti, deve necessariamente corrispondere un aumento della dimensione campionaria (per garantire una maggiore affidabilità delle stime) e viceversa.

  - La prima tabella (tabella 4) " ***Dimensione campionaria complessiva e sua sensibilità alla variazione del 10% della precisione desiderata***" fornisce la dimensione campionaria complessiva richiesta in base ai parametri impostati, e quindi alle variabili di input e ai vincoli indicati dall'utente. In corrispondenza a tale dato viene indicata la sensibilità del campione a fronte di una variazione del 10% della precisione desiderata.
  - La seconda (tabella 5) "***Sensibilità della dimensione del campione per tipo di dominio e variabile***" in particolare mostra poi come la sensibilità relativa all'intero campione viene ripartita all'interno di ciascun tipo di dominio e in base alle variabili considerate.

- La terza ed ultima sezione comprende le tabelle riportate in seguito:
  - La tabella "***Coefficienti di variazione programmati per tipo di dominio e variabile***", (tabella 6), richiama i vincoli scelti in precedenza (quelli del data-set originale o quelli relativi alle successive elaborazioni, se si è scelto di modificarli): riporta una riga per ciascun tipo di dominio che visualizza i coefficienti di variazione riferiti a tutte le variabili oggetto di studio.

- Viene poi stampata una tabella per ciascuna variabile (tabelle 7), "**Valori minimo medio e massimo dei coefficienti di variazione attesi della variabile per ciascun tipo di dominio**", in cui si riportano i tipi di dominio, il coefficiente di variazione stimato e gli estremi dell'intervallo di confidenza.
- Le ultime tabelle (tabelle 8), "**Coefficienti di variazione attesi in ciascun dominio e per ciascuna variabile per il tipo di dominio**", specificano i coefficienti di variazione attesi per ciascuna variabile ad un ulteriore livello di dettaglio, in quanto si riferiscono a tutte le modalità (i domini veri e propri) di ciascuno dei tipi di dominio.

Le stampe appena descritte sono richiamabili da entrambe le funzioni principali:

Si ottengono tramite la "**Visualizzazione Stampe**", nella schermata principale del software (figura 1 oppure figura 2, prima sezione), specificando il nome del data-set degli strati di riferimento e richiamando l'elaborazione desiderata (se è in memoria più di una elaborazione).

Come precedentemente specificato possono essere richiamate anche attraverso l'opzione "**Stampe**" relativa alla funzione principale "Selezione dei parametri" (si veda figura 2).

Parte II

Come premesso inizialmente nel documento, il progetto si struttura in due fasi: la prima fase progettuale ha portato alla realizzazione del software MAUSS, per il calcolo dell'allocazione ottimale nel caso multivariato e per più domini di stima in indagini ad un unico stadio; la seconda fase è stata pianificata per non trascurare le esigenze degli utenti che si occupano di indagini con disegni campionari a due stadi, nella fattispecie indagini sulle famiglie. La presente parte del documento si occupa della trattazione di questa seconda fase.

Analogamente a quanto fatto per la Parte I, viene dapprima introdotta la problematica generale dei disegni campionari a due stadi e vengono descritti alcuni cenni metodologici alla base di un nuovo modulo che andrà a costituire il nucleo elaborativo di una nuova funzione da implementare in MAUSS (paragrafo 4). Nel paragrafo 5 vengono descritti i passi procedurali alla base del nuovo modulo da integrare in MAUSS e nel paragrafo 6 vengono infine illustrati l'input e i dettagli procedurali per lo sviluppo di una nuova funzione in MAUSS.

4. Il problema dei disegni campionari a due stadi e lo studio di una nuova funzione in MAUSS di supporto per le indagini sulle famiglie

Il campo di applicazione individuato attualmente dal software per il caso di un disegno ad un solo stadio stratificato appare troppo limitato per le esigenze applicative delle indagini Istat.

Per questo motivo è stato avviato uno studio progettuale che si occupa del software in un contesto di indagini a due stadi.

Dallo studio è emerso che per risolvere la questione dei disegni a due stadi è possibile integrare il software MAUSS con una **funzione aggiuntiva**, alla cui base si pone un **modulo** di correzione dei risultati dell'algoritmo implementato per i disegni ad uno stadio, in modo da valutare anche l'effetto del disegno campionario a due stadi.

Tale aggiustamento – come si vedrà nel seguito - avviene in virtù della considerazione che, per i disegni a due stadi, si può tenere conto dell'eventuale peggioramento nella variabilità delle stime nella fase di definizione dei valori di input.

Allo stato attuale è stato costruito un primo prototipo software del nuovo modulo, illustrato in dettaglio nei prossimi paragrafi.

La progettazione del prototipo è stata eseguita basandosi sul lavoro di progettazione del disegno campionario per l'indagine Eusilc, ad opera di Stefano Falorsi e Claudia De Vitiis, e fondato a sua

volta sulla metodologia presentata da Falorsi e Russo (2002), che ben si adatta alle indagini sulle famiglie. In effetti l'estensione funzionale di MAUSS, seppure con carattere di generalità, è appropriata per risolvere la questione relativa alle indagini Istat sulle famiglie e in questo documento si riporta quanto analizzato e sviluppato a tale proposito.

Per condurre a termine la seconda fase progettuale e realizzare la nuova funzione per i disegni a due stadi, nel prossimo futuro si dovrà procedere con le seguenti attività:

1. Approfondimento di alcuni dettagli, che permetterebbero di considerare conclusa l'implementazione del prototipo software del nuovo modulo (ad esempio – come si vedrà nel seguito - nell'uso della procedura iterativa non è stato definito un criterio ottimale di convergenza);
2. Verifica generale del nuovo modulo implementato;
3. Definizione completa della estensione funzionale del software MAUSS - a partire dal nuovo modulo implementato – comprensiva di una nuova interfaccia utente.
4. Verifica sulla generalizzazione del caso qui analizzato, relativo alle indagini Istat sulle famiglie.

4.1 Alcune definizioni di base

Prima di procedere con la descrizione del nuovo modulo è bene chiarire alcuni concetti fondamentali. Nel seguito si fa infatti uso dei termini *cluster*, *grappolo* e *strato* che potrebbero essere interpretati in modo diverso da quello qui utilizzato, creando confusione. E' da osservare che per MAUSS gli *strati* sono partizioni di interesse alla base dei domini di stima e che, a tale proposito, è già stato scritta qualche specificazione nel precedente paragrafo 1. Per quanto riguarda invece il *cluster* e il *grappolo* è bene aggiungere qualche informazione:

- Cluster: si utilizza il concetto di cluster per non utilizzare il concetto di unità primaria, in quanto in molte indagini Istat sulle famiglie si fa spesso uso di un disegno di tipo complesso o misto. Tale disegno si avvale di due differenti schemi di campionamento, in quanto i comuni sono suddivisi in due sottoinsiemi sulla base della popolazione residente, e sempre sulla base della popolazione si definisce la stratificazione.

  Si distinguono:

  - l'insieme dei comuni Auto Rappresentativi (che indicheremo d'ora innanzi come comuni AR), formato dai comuni di maggiore dimensione demografica (per i quali la popolazione supera il valore di una certa soglia);
  - l'insieme dei comuni Non Auto Rappresentativi (o NAR), costituito dai rimanenti comuni.

  Solo nell'ambito dei comuni NAR viene adottato un disegno a due stadi con stratificazione delle unità primarie; le unità primarie sono costituite dai comuni e le unità secondarie dalle famiglie anagrafiche. Nell'ambito dell'insieme dei comuni AR, ciascun comune viene invece considerato come uno strato a sè stante; il disegno è ad uno stadio e le famiglie anagrafiche rappresentano unità primarie (che corrispondono alle unità finali, essendo il disegno ad un unico stadio).

  Per entrambi i disegni l'aggregazione di riferimento per il problema nel seguito trattato è la popolazione del comune, che rappresenta perciò il cluster di riferimento.

- Grappolo: generalmente le indagini sulle famiglie utilizzano sia per l'insieme dei comuni AR che NAR, il campionamento (ad uno o due stadi) a grappolo, dove il grappolo è la famiglia che costituisce un insieme di individui, tutti selezionati.

4.2 Il modulo alla base della nuova funzione da implementare in MAUSS

Il prototipo software del nuovo modulo, come già specificato, è stato implementato per correggere i risultati dell'algoritmo attualmente alla base di MAUSS, che calcola l'allocazione multivariata in campioni ad uno stadio, in modo da considerare l'eventuale influenza del disegno a due stadi.

E'utile evidenziare fin d'ora che nella pratica l'aggiustamento si realizza per mezzo di iterazioni sul calcolo dell'allocazione campionaria, modificando dunque il valore ottenuto ad ogni ciclo.

Le informazioni di input, nella fattispecie gli scarti quadratici medi delle variabili (S1, S2, …SP in figura 3), che spiegano la variabilità del fenomeno che si analizza, vengono corrette per mezzo di una statistica – il *deff* – che esprime l'effetto del disegno di campionamento.

Il *deff* è una misura appropriata a valutare l'effetto del disegno a due stadi rispetto al caso di un disegno casuale semplice in quanto:

- è dato dal rapporto tra la varianza del disegno di interesse e quella del disegno casuale semplice
- è esprimibile tramite una approssimazione in funzione dei coefficienti di correlazione intraclasse:

Considerando i valori per strato, si ha:

$$Deff=1+rho\ (b-1) \qquad (2)$$

dove con *b* si intende *il numero medio di osservazioni* negli strati e con *rho* si intendono i coefficienti di correlazione intraclasse. La radice quadrata del *Deff* corrisponde al cosiddetto *Deft* (esso viene utilizzato in particolare nel paragrafo successivo).

Sulla base della (2), per calcolare il *deff* e correggere l'allocazione, si deve partire dal calcolo del *b*. Dal momento che *b* è il numero medio di osservazioni, dipende dalla allocazione delle unità; di conseguenza si ha bisogno di una allocazione di partenza.

Per problemi ricorsivi di tale genere la soluzione più naturale consiste appunto nell'uso di una procedura iterativa.

E' infine da evidenziare che si è pensato di affinare la precisione di tale correzione tenendo conto di un diverso aggiustamento per i comuni autorappresentativi rispetto a quelli non

autorappresentativi nel modo seguente:

siano:

R = $(popolazione\ totale\ nel\ campione)\ /\ (popolazione\ totale\ nello\ strato)^2$

Rar = $(popolazione\ AR\ campionaria\ nello\ strato)\ /\ (popolazione\ AR\ nello\ strato)^2$

$Rnar$ = $(popolazione\ NAR\ campionaria\ nello\ strato)\ /\ (popolazione\ NAR\ nello\ strato)^2$

bdis_nar e bdis_ar = *numeri medi di osservazioni ottenuti considerando separatamente le parti AR e NAR*

rho_nar e rho_ar = *rho ottenuto considerando separatamente le parti AR e NAR*

Si ha rispettivamente per la parte NAR e AR:

$$Deff\_nar = 1+rho\_nar\ *(bdis\_nar-1) \qquad (3)$$

$$Deff\_ar = R*(\ 1/Rar*(1+rho\_ar*(bdis\_ar-1))+\ 1/Rnar\ *(1+rho\_nar*(bdis\_nar-1))\ ) \qquad (4)$$

La (2) è quindi esprimibile per il caso **autorappresentativo** e **non autorappresentativo** secondo le (3) e (4).

Il nuovo modulo di MAUSS comprenderà una prima fase di calcolo della *soglia* in base alla quale distinguere i comuni tra autorappresentativi e non autorappresentativi. Successivamente quindi procederà nel calcolo dei *deff* - secondo la (2) o la (3) per la parte NAR o AR - e di conseguenza della diversa correzione nella variabilità.

I passi e i dettagli procedurali delle fasi appena menzionati sono trattati nei prossimi paragrafi.

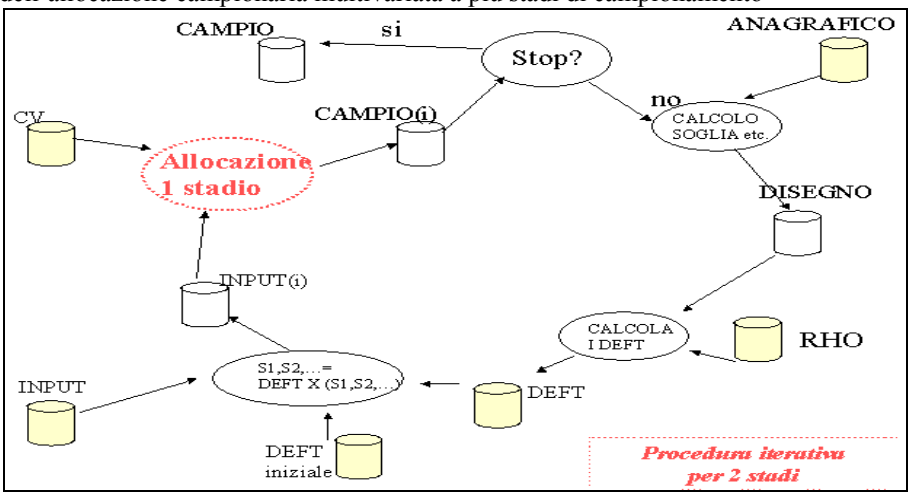
5. I passi procedurali alla base del nuovo modulo da integrare in MAUSS

In questo paragrafo e, con maggior dettaglio nel prossimo, vengono presentate le procedure che sono alla base dell'implementazione del nuovo modulo per il calcolo della allocazione campionaria nel caso multivariato per i disegni a due stadi.

Viene considerato il caso delle indagini Istat sulle famiglie introdotto nel paragrafo 4.1, con un disegno che si avvale di due differenti schemi di campionamento per i comuni autorappresentativi e non autorappresentativi.

I passi logici sottostanti la procedura vengono sintetizzati allo scopo di rappresentare in maniera comprensibile il flusso delle informazioni; nel paragrafo 6.2 le informazioni vengono invece dettagliate in forma di passi procedurali in linguaggio SAS.

Figura 11: la procedura iterativa per adattare la procedura di calcolo dell'allocazione campionaria multivariata a più stadi di campionamento



Nello sviluppo del nuovo modulo si sono considerati i seguenti passi:

Passo 1. Il primo passo corrisponde al calcolo del valore di una cosiddetta **soglia**, che permette di discriminare fra gli strati autorappresentativi e non autorappresentativi.

Come è possibile vedere dalla figura 11, per calcolare la soglia è necessario avere a disposizione una prima allocazione (CAMPIO(i) per i=1 in figura 11) e un data-set di dati anagrafici, riferito al *cluster* considerato (ANAGRAFICO in figura 11): l'**allocazione di partenza** si può ottenere tramite **un primo ciclo elaborativo**, in cui si utilizza l'algoritmo alla base del calcolo della allocazione che attualmente è già sviluppato in MAUSS e che alloca le unità negli strati a prescindere da qualsiasi aggiustamento (come si vedrà nel seguito, nell'iterazione non si considera il deft o, più precisamente, si pone deft=1); l'**anagrafico** deve essere disponibile come informazione di input (anagrafico *comunale*).

Passo 2. Sulla base della soglia si definiscono gli strati **autorappresentativi** e **non autorappresentativi**

Passo 3. Si passa al calcolo dei **deft**, a partire dai deff. Dalle formule (2) (3) e (4) appare chiaro che si deve disporre, oltre all'**anagrafico** di cui sopra, dei valori dei **rho,** eventualmente distinti tra strati autorappresentativi e non autorappresentativi (RHO in figura 11).
(Come si evincerà nel seguito nel caso multivariato, esistono p variabili, e dunque si considera un deff per ciascuna variabile: deff1, deff2, … deffp; per i rho inoltre i valori raddoppiano per considerare entrambi i casi AR e NAR).

Passo 4. Sulla base dei **rho** e delle popolazioni desumibili dall'**anagrafico**, si calcolano **nuovi deff** (e di conseguenza i deft), che tengono anche conto della ripartizione degli strati in autorappresentativi e non (in figura 11 i passi 3 e 4 sono schematizzati nel processo "*calcola i deft*").

Passo 5. I nuovi **deft** entrano in input in un **secondo ciclo**, in quanto si moltiplicano i valori degli scarti quadratici medi delle variabili nella popolazione (S1, S2, …SP) per i deft (s1=s1*deft1, s2=s2*deft2, …. sp=sp*deftp).

Applicando nuovamente il software ai dati di input in cui sono state modificate le variabili S1, S2 ….SP, si ottiene una nuova allocazione, *una **nuova soglia**, **nuove popolazioni** da moltiplicare per i **rho*** e così via.

Il procedimento viene iterato fino al soddisfacimento di un criterio di convergenza, che ferma l'algoritmo. Tale criterio è ancora da definire e si basa sull'analisi empirica, in quanto nelle applicazioni reali si è osservata una tendenza al raggiungimento di una stabilità nei risultati dopo qualche iterazione.

6. L'input e i dettagli procedurali per lo sviluppo di una nuova funzione in MAUSS

Per procedere a livello di dettaglio espositivo maggiore, in questo paragrafo si fa uso di una terminologia strettamente legata al linguaggio SAS: si descrivono le informazioni di input necessarie per il funzionamento del prototipo software implementato in termini di data-set SAS; allo stesso modo si riportano i dettagli procedurali in termini di programmazione SAS.

6.1 I data-set di input della funzione per i due stadi

Il nuovo modulo necessita di informazioni di input addizionali rispetto a quelle richieste dal funzionamento di MAUSS per i disegni ad uno stadio. Sono necessari infatti in totale cinque data-set di partenza: i due data-set già descritti nel paragrafo 2.1 (siano **INPUT** – denominazione del data-set degli strati - e **VINCOLI** -denominazione del data-set dei vincoli) e tre aggiuntivi (**DEFT**, **ANAGRAFICO** e **RHO)**.
Questi ultimi vengono qui illustrati.

Supponendo di avere rilevato p variabili, si ottiene:

(1) data-set dei deft: **Dataset DEFT**

Variabili:
STRATO = strato
DEFT1 =deft - prima variabile
……
DEFTP =deft – p-ma variabile

Strato	Deft1		Deftp
1		…….	
2		…….	
……			
…….			

(2) data-set delle popolazioni: **Dataset ANAGRAFICO** (per strato e cluster)

Variabili:
STRATO =strato
CLU =Codice cluster considerato (comune)
POP =Popolazione nell'universo del cluster (nel campione a due stadi, con riferimento alle unità finali, si hanno gli individui - se il campionamento è a grappoli - o le famiglie se se non è a grappoli)
MINIMO =nro minimo di interviste per strato
DELTA =nro medio di individui nella famiglia (grappolo)

STRATO	CLU Cluster (Comune)	POP Unità Elem. (Individui)	MINIMO (per strato o anche unico)	DELTA (per strato o anche unico)
1	Comune 1	10.000		
1	Comune 2	15.000		
1	Comune 3	5.000		
1	Comune 4	3.000		
2				
……..				

(3) data-set dei rho: **Data-set RHO**

Variabili:
STRATO
RHO-ar1
RHO-nar1
……
RHO-arp
RHO-narp

Strato	RHO-ar1	RHO-nar1		RHO-arp	RHO-narp
1			…….		
2			…….		
……					
….					

6.2 I passi procedurali per la correzione della allocazione campionaria per disegni a due stadi sulle famiglie (in sas language):

Per semplificare la successiva esposizione, si suppone che ci siano solo 2 variabili.

dataset di input: **INPUT, DEFT, ANAGRAFICO, RHO**
dataset di output: **DISEGNO**

Si parte da un primo ciclo in cui il dataset dei deft, non avendo informazioni utili – viene definito con tutti i valori pari all'unità.

****PRIMO CICLO****

Passo 1

Nel data-set DEFT - che ha per variabili STRATO DEFT1 DEFT2 - si pone deft1=1 e deft2 =1 per ciascuno strato:

Strato	Deft1	Deft2
1	1	1
2	1	1
….	……	……..
….	1	1

****DAL SECONDO CICLO IN POI****

Passo 2

step SAS – passo 2

```
data beth.INPUT; merge beth.deft beth.INPUT; BY STRATO;
keep strato c cens dom1-dom3 m1-m2 s1-s2 n;
s1=s1*deft1;
s2=s2*deft2;
```

COMMENTO al passo 2
Si modifica il data-set INPUT (il data-set degli strati descritto nel paragrafo 2.1) tramite i deft, ovvero tramite i valori del data-set DEFT.

Passo 3

Si applica MAUSS e si ottiene la variabile **CAMP** che viene memorizzata nel data-set **CAMPIO** (che ha le variabili Strato e Camp)

Passi 4, 5, 6

step SAS – passi 4, 5

```
PROC SUMMARY DATA=BETH.ANAGRAFICO NWAY; CLASS strato;
OUTPUT OUT=popol
SUM (POP)= POP_str;                    *** dataset temporaneo POPOL ha strato e pop_str;



 Data UNO; merge BETH.ANAGRAFICO BETH.CAMPIO POPOL;
 by strato;
   F =   CAMP/Pop_str;              * F nello strato;
  SOGLIA   = (MINIMO /  F )* DELTA;
 AR   = 0;
 IF (pop >=SOGLIA  ) THEN AR  =1;  /* si confronta la popolazione del cluster
                                      con la soglia, calcolata a livello di strato */


 POPAR =POP*AR  ;                  * popar=individui popolazione ar – la variabile è pari a 0 o alla
popolazione del cluster;
```

step SAS – passo 6

```
****creazione di data-set DISEGNO ***;

PROC SUMMARY DATA=UNO NWAY; CLASS strato;
id pop_str minimo delta soglia f camp:
OUTPUT OUT =A
SUM(POPAR AR)= popar num_ar;

DATA A; SET A;  drop _type_ _freq_
  num_cl= _freq_;
  num_nar=num_cl-num_ar;
  popnar=pop_str-popar;
  camp_ar=round(f*popar);
  campnar=camp-camp_ar;

DATA BETH.DISEGNO; SET A;
KEEP strato   pop_str   popar   popnar   num   num_cl   num_ar   num_nar
camp   camp _ar   camp _nar minimo delta soglia f;
```

COMMENTO ai Passi 4, 5, 6

Commento al Passo 4

- Si calcola la **soglia** per definire gli strati autorappresentativi e non autorappresentativi .

I dati di input per questo passo derivano dal data-set ANAGRAFICO (variabile **pop_str**, ottenuta come somma di pop) e dall'applicazione del software per l'allocazione ad uno stadio (variabile **camp** del data-set CAMPIO).

a) Calcolo la frazione di campionamento per strato
F nello strato = ***camp / pop_*****str**
b) Calcolo la soglia:
soglia nello strato = ***(minimo/F) x delta***

Commento al Passo 5
- Sulla base della soglia si definisce lo strato **autorappresentativo** o **non autorappresentativo**.

Commento al Passo 6
Si calcolano una serie di informazioni che vanno nel data-set **DISEGNO** di output, aggregato per strato.

Approfondimenti sulle informazioni contenute nel data-set ANAGRAFICO .

Nel data-set sono richieste le popolazioni dei comuni, in quanto in Istat spesso le indagini sulle famiglie utilizzano l'archivio dei dati anagrafici comunali.

L'algoritmo implementato nel nuovo modulo necessita invece delle informazioni a livello di strato, pertanto al suo interno viene effettuato un passaggio di aggregazione delle popolazioni (variabile *POP*) dei comuni, per determinare la popolazione degli strati di interesse.

Se il disegno a due stadi è a grappoli, esistono sia le unità finali (le famiglie) che le unità elementari (*POP*), ma l'algoritmo necessita solo della popolazione dello strato in termini di unità elementari.

Se la famiglia corrisponde ad un grappolo, per DELTA si intende il numero medio di individui per famiglia (ad esempio in Italia risulta attualmente pari a 2.8 approssimativamente, ma può variare per regione/strato) e per MINIMO il numero di interviste per strato.

Tra le variabili del data-set ANAGRAFICO si è pensato di inserire anche la variabile DELTA (che eventualmente si pone uguale ad 1), in quanto molte indagini Istat sulle famiglie dispongono dei dati relativi alla variabile MINIMO (numero di interviste per strato) espressi in termini di famiglie (il grappolo). In altri termini si considera il numero di famiglie da intervistare e non la popolazione. Nel calcolo della soglia occorre considerare il numero di individui da intervistare, e dunque nell'algoritmo il DELTA è utile come moltiplicatore (come sopra accennato per le indagini in cui ciò non risulta utile, nel data-set basta porre tutti i valori della variabile pari ad 1).

Una ultima osservazione riguarda il fatto che nel calcolo della soglia si distingue tra popolazione di strati autorappresentativi e non, ma gli strati sono formati da diversi comuni. In questo modo, uno strato autorappresentativo può contenere diversi comuni autorappresentativi, mentre – secondo quanto sopra esposto (paragrafo 4.1) - nella effettiva stratificazione un comune autorappresentativo costituirà uno strato a se stante.

Passo 7

Step SAS – passo 7

```
data d; merge BETH.disegno  BETH.rho BETH.DEFT;
 by strato;
bdis_ar=1*delta;
bdis_nar=minimo*delta;
```

```
array rho_ar(2) rho_ar1-rho_ar2;
array rho_nar(2) rho_nar1-rho_nar2;

array deff(2) deff1-deff2;
array deft(2) deft1-deft2;

 do i=1 to 2;

  deff(i)= 1+rho_nar(i)*(bdis_nar-1) ;   ****NAR;

  if popar ne 0 then
   deff(i)=(camp/pop_str**2)*((popar**2/camp_ar)*(1+rho_ar(i)*(bdis_ar-1))+
        (popnar**2/camp_nar)*(1+rho_nar(i)*(bdis_nar-1)));
  if deff(i)<0 then deff(i)=1;                              ***AR;
  deft(i)=sqrt(deff(i));
 end;
proc sort; by strato;
run;
```

COMMENTO al Passo 7
Sulla base di DISEGNO e di RHO, si calcolano i **nuovi deft**

Passo 8

Step sas - Passo 8

```
data s; merge d BETH.INPUT;
keep strato c cens dom1-dom3 m1-m2 s1-s2 n deft1-deft2 ;
s1=s1*deft1;
s2=s2*deft2;

data BETH.input2; set d;  * data-set memorizzato;
data BETH.deft2; set d;
keep strato deft1-deft2 ;  * solo i deft nuovi per ogni ciclo;
```

COMMENTO al Passo 8
Si moltiplicano i nuovi deft per le S*i* iniziali:

Il nuovo deff entra in input in un **secondo ciclo**: deff=nuovo valore

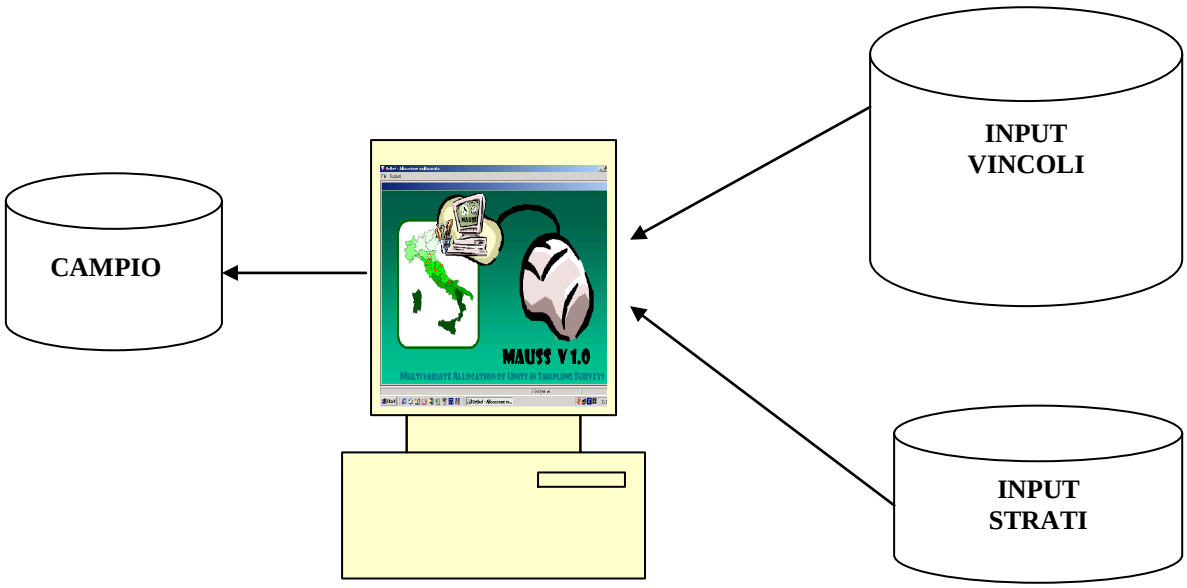
- SI TORNA AL PASSO 2

Parte III

LA DOCUMENTAZIONE PROGETTUALE

Nella Parte III sono raccolti i documenti progettuali relativi allo sviluppo del software MAUSS.
La Parte III comprende anche del materiale progettuale tecnico, allo scopo di stendere un registro consultabile, in cui si elenca la lista dei dataset e delle variabili della procedura e la descrizione dei moduli.

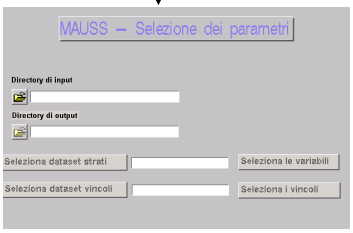
SOFTWARE PER ALLOCAZIONE CAMPIONARIA



Sequenza pannelli.

Stampe

File → Scelta progetto
Funzioni → Selezione parametri



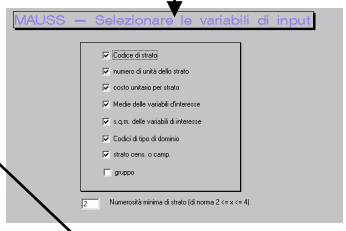
Seleziona variabili
Dataset strati

Seleziona vincoli

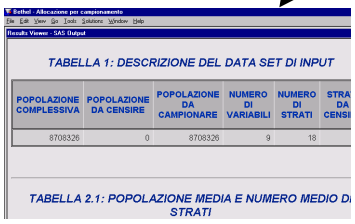
Funzioni → Stampe

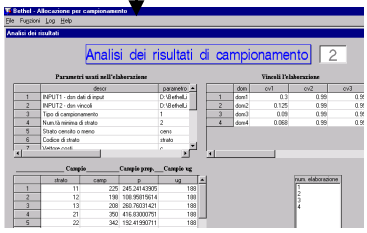
Esegui Procedura

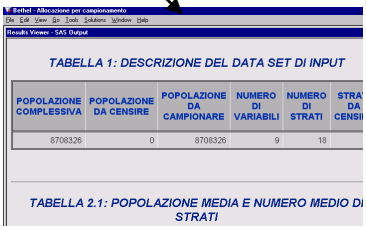
Analisi della popolazione → Dimensione campione



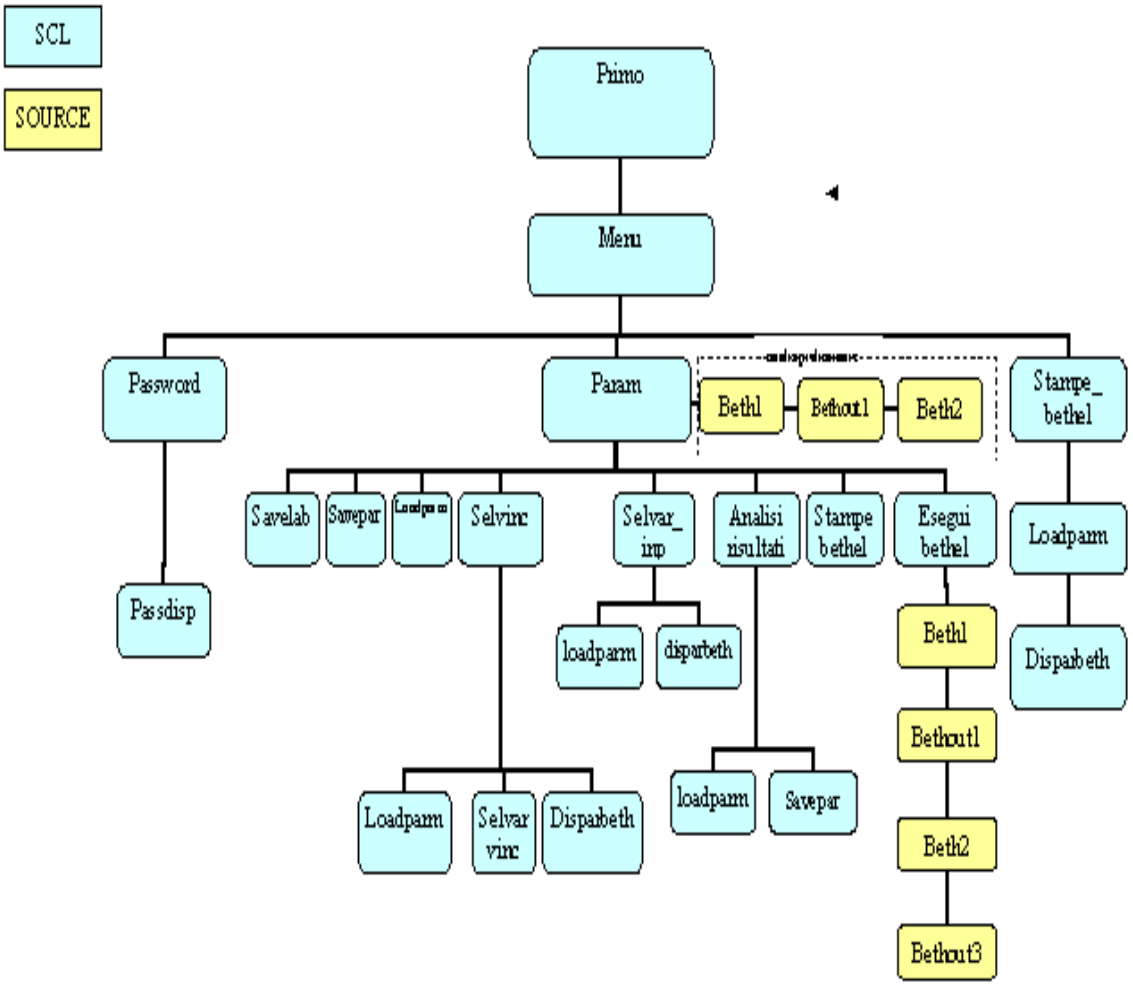
Analisi dei Risultati campionamento



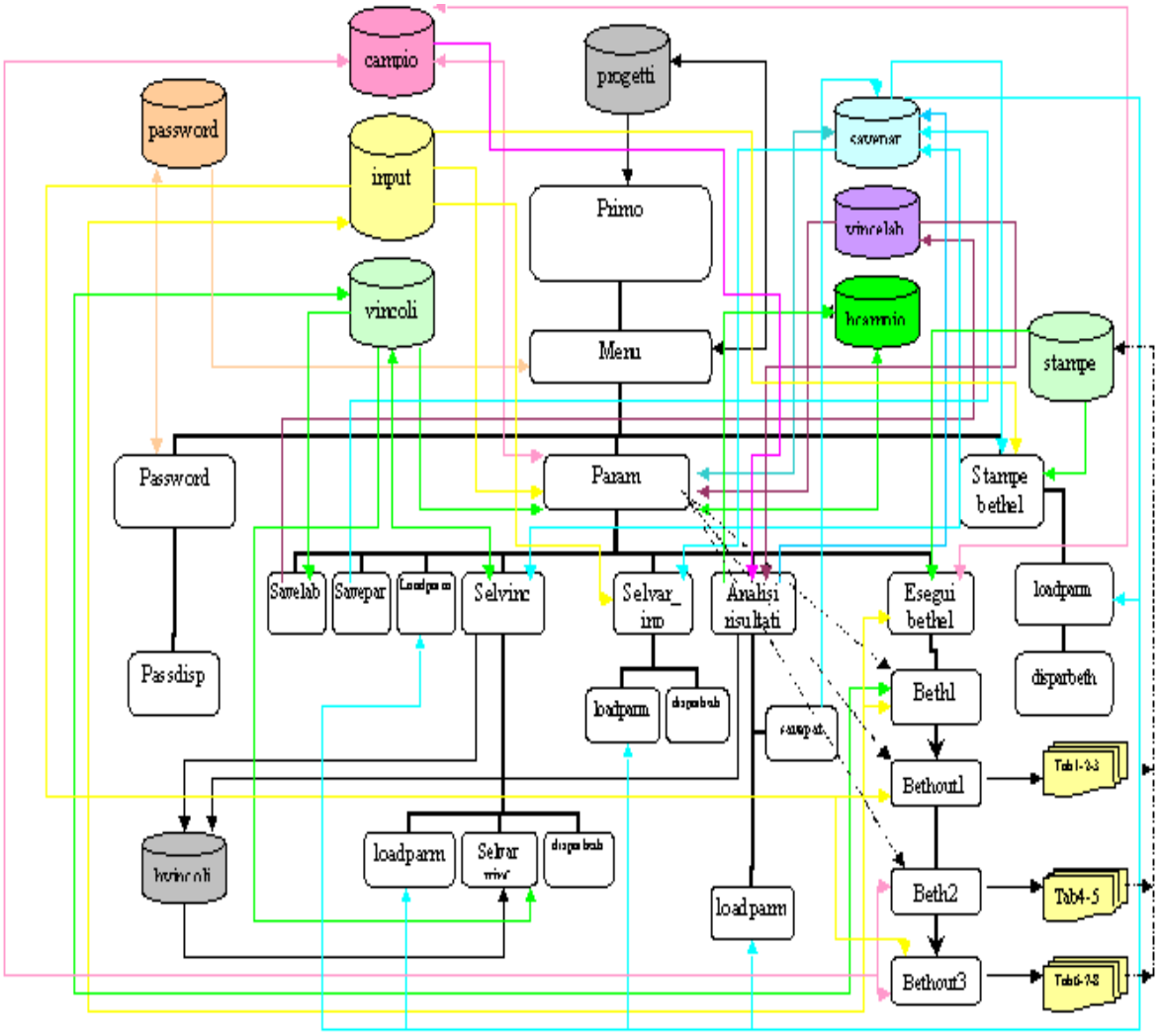




Sequenza moduli sviluppati



Flusso dei dati



REGISTRO DATASET E VARIABILI

INPUT
STRATI

column name	type	length	format	descrizione
Strato	number	8		Codice dello strato
N	number	8		Numero delle unità della popolazione
dom1	number	8		Tipo di dominio 1
dom2	number	8		Tipo di dominio 2
dom3	number	8		Tipo di dominio 3
dom4	number	8		Tipo di dominio 4
m1	number	8		Media della prima variabile nella popolazione
m2	number	8		Media della seconda variabile nella popolazione
m3	number	8		Media della terza variabile nella popolazione
m4	number	8		Media della quarta variabile nella popolazione
m5	number	8		Media della quinta variabile nella popolazione
m6	number	8		Media della sesta variabile nella popolazione
m7	number	8		Media della settima variabile nella popolazione
m8	number	8		Media della ottava variabile nella popolazione
m9	number	8		Media della nona variabile nella popolazione
s1	number	8		S.q.m. della prima variabile nella popolazione
s2	number	8		S.q.m. della seconda variabile nella popolazione
s3	number	8		S.q.m. della terza variabile nella popolazione
s4	number	8		S.q.m. della quarta variabile nella popolazione
s5	number	8		S.q.m. della quinta variabile nella popolazione
s6	number	8		S.q.m. della sesta variabile nella popolazione
s7	number	8		S.q.m. della settima variabile nella popolazione
s8	number	8		S.q.m. della ottava variabile nella popolazione
s9	number	8		S.q.m. della nona variabile nella popolazione
C	number	8		Costo
Cens	number	8		Variabile indicatrice dello strato da censire o campionare
popcens	*number*	*8*		Popolazione censuaria
Camp	number	8		Allocazione campionaria di Bethel (scritto in output, al termine della procedura)

INPUT
VINCOLI

column name	type	length	format	descrizione
dom	text	4	$4.	Codice di dominio
cv1	number	8	best8.	Coeff. di variazione della prima variabile
cv2	number	8	best8.	Coeff. di variazione della seconda variabile
cv3	number	8	best8.	Coeff. di variazione della terza variabile
cv4	number	8	best8.	Coeff. di variazione della quarta variabile
cv5	number	8	best8.	Coeff. di variazione della quinta variabile
cv6	number	8	best8.	Coeff. di variazione della sesta variabile
cv7	number	8	best8.	Coeff. di variazione della settima variabile
cv8	number	8	best8.	Coeff. di variazione della ottava variabile
cv9	number	8	best8.	Coeff. di variazione della nona variabile

CAMPIO

column name	type	length	format	descrizione
strato	number	8		Codice di strato
camp	number	8		Allocazione campionaria di Bethel

PROGETTI

column name	type	length	format	descrizione
pathin	text	200		Percorso della directory di input
pathout	text	200		Percorso della directory di output
dsninp	text	200		Data-set di input
dsnvin	text	200		Data-set di output

HCAMPIO

column name	type	length	format	descrizione
nelab	number	8		Numero di elaborazione effettuata
strato	number	8		Codice di strato
camp	number	8		Allocazione campionaria di Bethel
p	number	8		Allocazione campionaria proporzionale
ug	number	8		Allocazione campionaria uguale

STRATA_SA
VEPAR

column name	type	length	format	descrizione
descr	text	35		Descrizione dei parametri utilizzati
parametro	text	2000		Parametri utilizzati per l'elaborazione

VINCELAB

column name	type	length	format	descrizione
nelab	number	8		Numero di elaborazione effettuata
dom	text	4	$4.	Codice di dominio
cv1	number	8		Coeff. di variazione della prima variabile
cv2	number	8		Coeff. di variazione della seconda variabile
cv3	number	8		Coeff. di variazione della terza variabile
cv4	number	8		Coeff. di variazione della quarta variabile
cv5	number	8		Coeff. di variazione della quinta variabile
cv6	number	8		Coeff. di variazione della sesta variabile
cv7	number	8		Coeff. di variazione della settima variabile
cv8	number	8		Coeff. di variazione della ottava variabile
cv9	number	8		Coeff. di variazione della nona variabile

HVINCOLI

column name	type	length	format	descrizione
nitera	number	8		Numero di iterazioni
dom	text	4	$4.	Codice di dominio
cv1	number	8		Coeff. di variazione della prima variabile
cv2	number	8		Coeff. di variazione della seconda variabile
cv3	number	8		Coeff. di variazione della terza variabile
cv4	number	8		Coeff. di variazione della quarta variabile
cv5	number	8		Coeff. di variazione della quinta variabile
cv6	number	8		Coeff. di variazione della sesta variabile
cv7	number	8		Coeff. di variazione della settima variabile
cv8	number	8		Coeff. di variazione della ottava variabile
cv9	number	8		Coeff. di variazione della nona variabile

PASSWORD

column name	type	length	format	descrizione
passw	text	15		Codice password

DESCRIZIONE DEI MODULI (nome libreria “bethet”)

MODULO: PRIMO **TIPO**: SCL (SENZA FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Prepara dati per menu generale.
Imposta il Pmenu che viene memorizzato nel catalogo BETHEL.SOURCE.
Viene invocato dal modulo di configurazione CFG con l’istruzione:
“-initcmd "af c=bethel.appl.primo.frame”

DS INPUT: BETHEL.PROGETTI (file dei progetti eseguiti)

CHIAMA: MENU.FRAME

MODULO: MENU **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Visualizza e gestisce il Menu generale.
Crea il dataset progetti se non esiste - se esiste lo aggiorna

DS INPUT: BETHEL.PASSWORD (file delle password di abilitazione)
BETHEL.PROGETTI (file dei progetti eseguiti)

DS OUTPUT: BETHEL.PROGETTI (file dei progetti eseguiti)

CHIAMATO DA: PRIMO.FRAME

CHIAMA: PASSWORD.SCL (controllo password di autorizzazione)
PARAM.FRAME (parametri passati: pathin, pathout, dsninp, dsnvin, da_progetto
STAMPE.FRAME ((parametri passati: pathout)

MODULO: PASSWORD **TIPO**: SCL (SENZA FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Gestisce la password di autorizzazione.

DS INPUT: BETHEL.PASSWORD (file delle password di abilitazione)

DS OUTPUT: BETHEL.PASSWORD (file delle password di abilitazione)

CHIAMATO DA: MENU.FRAME

CHIAMA: PASSDISP.FRAME (parametri passati: cods, passw)

MODULO: PASSDISP **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Visualizza richiesta password di autorizzazione.

DS INPUT:

DS OUTPUT:

CHIAMATO DA: PASSWORD.FRAME

CHIAMA:

MODULO: PARAM **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Consente la selezione delle Directory di input e di output; dei dataset INPUT e VINCOLI e delle relative variabili e vincoli. Esegue, a richiesta, l'analisi preliminare che visualizza le tavole 1-2-3-4-5.

DS INPUT: BETHEL.STRATA (file di input)
BETHEL.SAVEPAR (parametri della elaborazione)
BETHEL.ERRORI (file dei vincoli)
BETHEL.VINCELAB (file dei vincoli per tutte le iterazioni)
BETHEL.CAMPIO (file della sensibilità del campionamento)
BETHEL.HCAMPIO (file campio contenente tutte le iterazioni)

DS OUTPUT: BETHEL.CAMPIO (file della sensibilità del campionamento)

CHIAMATO DA: MENU.FRAME

CHIAMA: STAMPE_BETHEL.FRAME (parametri passati: ,strata, errori, pathin, pathout, lparam swint, swvinc, numvar, pps, minstr, tpstr, strato, vetcos, vetn, int ,dominio, sqm, gruppo, vincoli, nitera, nelab, hvincoli)

ANALISI_RISULTATI.FRAME (parametri passati: strata, errori, pathin, pathout, pps, minstr, tpstr, strato, vetcos, vetn, int, hvincoli, dominio, sqm,gruppo, sw, cv, nitera, nelab, swelab)

SELVAR_INP.FRAME (parametri passati: strata, errori, pathin, pathout, lparam, swint, swvinc, numvar, pps, minstr, tpstr, strato, vetcos, vetn, int,dominio, sqm, gruppo, vincoli, cv, nitera, nelab, hvincoli)

SELVINC.FRAME (parametri passati: strata, errori, pathin, pathout, lparam swint, swvinc, numvar, pps, minstr, tpstr, strato, vetcos, vetn, int, dominio, sqm, gruppo, vincoli, cv, nitera, hvincoli, nelab)

LOADPARM.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr, tpstr, strato, vetcos, vetn, int, hvincoli, dominio, sqm, gruppo, sw, cv, nitera, nelab)

SAVEPAR.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr, tpstr, strato, vetcos, vetn, int, hvincoli, dominio, sqm, gruppo, sw , cv , nitera, nelab)

SAVELAB.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr, tpstr, strato, vetcos, vetn, int, hvincoli, dominio, sqm, gruppo, sw, cv, nitera, nelab)

ESEGUI_BETHEL.SCL (parametri passati: strata, errori, libinp, libout, lparam, swint, swvinc, pps, minstr, int, dominio, gruppo, vincoli, campio, nitera, hvincoli)

BETHEL.SOURCE.BETH1.SOURCE
BETHEL.SOURCE.BETHOUT1.SOURCE
BETHEL.SOURCE.BETH2.SOURCE

MODULO: SAVELAB **TIPO**: SCL (SENZA FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Memorizza i vincoli usati per l'elaborazione.

DS INPUT: BETHEL.ERRORI (file dei vincoli)

DS OUTPUT: BETHEL.VINCELAB (file dei vincoli per tutte le iterazioni)

CHIAMATO DA: PARAM.FRAME

CHIAMA:

MODULO: LOADPARM **TIPO**: SCL (SENZA FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Legge i parametri di una precedente elaborazione.

DS INPUT: BETHEL.SAVEPAR (file dei parametri della elaborazione)

DS OUTPUT

CHIAMATO DA: PARAM.FRAME
SELVINC.FRAME
SELVAR_INP.FRAME
ANALISI_RISULTATI.FRAME
STAMPE_BETHEL.FRAME

CHIAMA:

MODULO: SAVEPAR **TIPO**: SCL (SENZA FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Memorizza i parametri dell'elaborazione.

DS INPUT:

DS OUTPUT: BETHEL.SAVEPAR (file dei parametri della elaborazione)

CHIAMATO DA: PARAM.FRAME
ANALISI_RISULTATI.FRAME

CHIAMA:

MODULO: SELVINC **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Consente la selezione o la modifica dei vincoli.

DS INPUT: BETHEL.ERRORI (file dei vincoli)
BETHEL.SAVEPAR (file dei parametri della elaborazione)

DS OUTPUT: BETHEL.ERRORI (file dei vincoli)
BETHEL.HVINCOLI (file dei vincoli modificati)
BETHEL.SAVEPAR (file dei parametri della elaborazione)

CHIAMATO DA: PARAM.FRAME

CHIAMA: DISPARBETH.FRAME

SELVAR_VINC.FRAME (parametri passati: strata, errori, pathin, pathout, lparam, swint, swvinc, numvar, pps, minstr, tpstr, strato, vetcos, vetn, int, dominio, sqm, gruppo, vincoli, cv, nitera, nelab, hvincoli)

LOADPARM.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr,tpstr, strato,vetcos, vetn, int, hvincoli,dominio, sqm , gruppo, sw , cv , nitera)

MODULO: SELVAR_VINC **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Gestione dei vincoli.

DS INPUT: BETHEL.ERRORI (file dei vincoli)
BETHEL.HVINCOLI (file dei vincoli modificati)

DS OUTPUT:.

CHIAMATO DA: SELVINC.FRAME

CHIAMA:

MODULO: DISPARBETH **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Visualizza parametri utilizzati.

DS INPUT:

DS OUTPUT:.

CHIAMATO DA: SELVINC.FRAME
SELVAR_INP.FRAME
STAMPE_BETHEL.FRAME

CHIAMA:

MODULO: SELVAR_INP **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Selezione delle variabili di input.

DS INPUT: BETHEL.STRATA (file di input)
BETHEL.SAVEPAR (file dei parametri della elaborazione)

DS OUTPUT:

CHIAMATO DA: PARAM.FRAME

CHIAMA: DISPARBETH.FRAME

LOADPARM.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr, tpstr,strato, vetcos, vetn, int, vincoli, dominio, sqm , gruppo, sw , cv , nitera, nelab)

MODULO: ANALISI_RISULTATI **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Visualizza, per ogni iterazione, i parametri, i vincoli e il dataset campio. Consente di richiamare ogni singola iterazione.

DS INPUT: BETHEL.VINCELAB (file dei vincoli per tutte le iterazioni)
BETHEL.HVINCOLI (file dei vincoli modificati)
BETHEL.CAMPIO (file della sensibilità del campionamento)

DS OUTPUT: BETHEL.HCAMPIO (file campio contenente tutte le iterazioni)

CHIAMATO DA: PARAM.FRAME

CHIAMA: LOADPARM.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr, cens, strato,c, n, int, vincoli,dominio, sqm , gruppo, sw , cv , nitera, nelab)

MODULO: STAMPE.BETHEL **TIPO**: SCL (CON FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Consente di selezionare le stampe di ogni iterazione.

DS INPUT: BETHEL.STAMPE_BETHEL (file delle stampe)

DS OUTPUT:

CHIAMATO DA: PARAM.FRAME

CHIAMA: LOADPARM.SCL (parametri passati: strata, errori, pathin, pathout, pps, minstr, tpstr, strato, vetcos, vetn, int, hvincoli, dominio, sqm, gruppo, sw, cv, nitera, nelab)

DISPARBETH.FRAME

MODULO: ESEGUI_BETHEL **TIPO**: SCL (SENZA FRAME)

CATALOGO: BETHEL.APPL

FUNZIONE: Lancia la esecuzione dei programmi di allocazione e visualizza tutte le stampe.

DS INPUT: BETHEL.STAMPE_BETHEL (file delle stampe)
BETHEL.STRATA (file di input)
BETHEL.CAMPIO (file della sensibilità del campionamento)

DS OUTPUT: BETHEL.CAMPIO (file della sensibilità del campionamento)

CHIAMATO DA: PARAM.FRAME

CHIAMA: BETH1.SOURCE
BETHOUT1.SOURCE
BETH2.SOURCE
BETHOUT3.SOURCE

MODULO: BETH1 **TIPO**: SOURCE

CATALOGO: BETHEL.SOURCE

FUNZIONE: Crea il dataset di lavoro.

DS INPUT: BETHEL.STRATA (file di input)
BETHEL.ERRORI (file dei vincoli)

DS OUTPUT: BETHEL.STRATA (file di input)
BETHEL.ERRORI (file dei vincoli)

CHIAMATO DA: ESEGUI_BETHEL.FRAME

CHIAMA:

MODULO: BETHOUT1 **TIPO**: SOURCE

CATALOGO: BETHEL.SOURCE

FUNZIONE: Stampa le tabelle 1 2 3.

DS INPUT: BETHEL.STRATA (file di input)

DS OUTPUT: BETHEL.STAMPE_BETHEL (file delle stampe)

CHIAMATO DA: ESEGUI_BETHEL.FRAME (viene eseguito sempre dopo BETH1)

CHIAMA:

MODULO: BETH2 **TIPO**: SOURCE

CATALOGO: BETHEL.SOURCE

FUNZIONE: Procedura IML per la determinazione del campione e l'analisi di sensibilità. Stampa le tabelle 4 e 5

DS INPUT: BETHEL.STRATA (file di input)
BETHEL.ERRORI (file dei vincoli)

DS OUTPUT: BETHEL.CAMPIO (file della sensibilità del campionamento)
BETHEL.STAMPE_BETHEL (file delle stampe)

CHIAMATO DA: ESEGUI_BETHEL.FRAME (se variabile “stampe” = 2 or 3)

CHIAMA:

MODULO: BETHOUT3 **TIPO**: SOURCE

CATALOGO: BETHEL.SOURCE

FUNZIONE: Stampa le tabelle 6 7 8.

DS INPUT: BETHEL.STRATA (file di input)
BETHEL.CAMPIO (file della sensibilità del campionamento)

DS OUTPUT: BETHEL.STAMPE_BETHEL (file delle stampe)

CHIAMATO DA: ESEGUI_BETHEL.FRAME (viene eseguito sempre dopo BETH1 e BETH2, (se variabile “stampe” = 2 or 3)

CHIAMA:

BIBLIOGRAFIA

BETHEL, J. (1989) , “Sample Allocation in Multivariate Survey”, *Survey Metho*dology, 15, pp. 47-57.

CHROMY, J. (1987), “Designe Optimization with Multiple Objectives”, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 194-199.

FALORSI P.D., BALLIN M., DE VITIIS C., SCEPI G. (1998), “Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sule imprese condotte dall’Istat “, Statistica Applicata, vol.10 n°2, pp. 235-257.

FALORSI S., RUSSO A. (2002), "Il disegno di rilevazione per le indagini panel sulle famiglie” in *Quaderni di ricerca, Rivista di Statistica Ufficiale,* n. 3/2001, ISTAT .