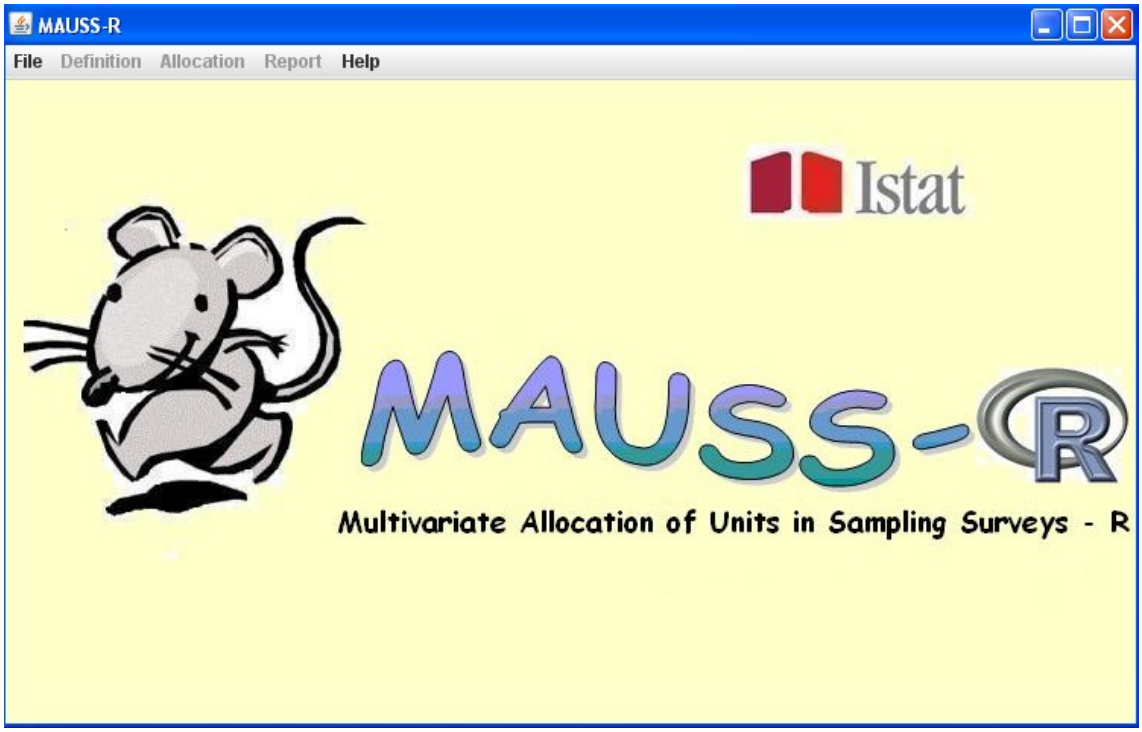
# MAUSS

## Multivariate Allocation of Units in Sampling Surveys




## Manuale utente e metodologico

Teresa Buglielli, Claudia De Vitiis, Giulio Barcaroli
*Servizio Metodi, Strumenti e Supporto metodologico*
*Direzione Tecnologie e Supporto Metodologico*
Istituto Nazionale di Statistica

# Indice generale

# Introduzione

MAUSS è uno strumento per la definizione del disegno di campionamento per le indagini campionarie da popolazioni finite. Garantisce al contempo criteri di ottimalità, flessibilità e facile gestione per i soggetti che hanno la responsabilità di progettare e condurre le tali indagini.

Consente all'utente, una volta definiti gli obiettivi ed i vincoli operativi dell'indagine che si vuole progettare, di scegliere il disegno di campionamento migliore tra quelli ottenibili adottando differenti definizioni delle caratteristiche fondamentali dell'indagine, quali, ad esempio, il tipo di stratificazione, la precisione desiderata delle stime, la dimensione del campione, la tipologia dei domini di studio, le variabili d'interesse.

L'utilizzo di tale software, inoltre, garantisce la trasparenza, la standardizzazione e la correttezza delle metodologie utilizzate.

L'attuale versione di MAUSS è una evoluzione di precedenti applicazioni, sviluppate in SAS. La progettazione e lo sviluppo di queste prime versioni è dovuta a metodologi ed informatici (tra i primi citiamo Marco Ballin, Claudia De Vitiis, Piero Demetrio Falorsi, Germana Scepi; tra i secondi ancora Marco Ballin e Piero Demetrio Falorsi, unitamente a Daniela Pagliuca, Paolo Floris e Roberto Di Giuseppe).

La decisione di migrare la versione SAS verso R è stata presa nel quadro di una strategia che tende a ridurre la dipendenza da software proprietari e a garantire la piena portabilità degli strumenti generalizzati sviluppati dall'ISTAT. Inoltre, sono state aggiunte nuove funzioni, all'interno di una interfaccia più avanzata. Lo sviluppo della versione descritta in questo manuale è dovuto a Teresa Buglielli (interfaccia Java per gestione progetti ed esecuzione moduli), Daniela Pagliuca (implementazione della metodologia in R) e Giulio Barcaroli (algoritmo di Chromy in Fortran).

# 1. La metodologia implementata in MAUSS

## *1.1. Definizione del problema metodologico dell'allocazione*

### 1.1.1. La progettazione del disegno campionario di un'indagine

Nella progettazione di un'indagine campionaria, la fase di studio del disegno di campionamento e della definizione della numerosità campionaria e della sua allocazione tra gli strati richiede la definizione di una serie di parametri e di informazioni, da cui deriva poi la costruzione dell'input per la procedura di allocazione. E' necessario infatti definire:

- la popolazione di interesse
- l'unità di campionamento,
- l'archivio di selezione contenente le unità della popolazione,
- le variabili di interesse,
- i parametri che costituiscono oggetto di stima,
- il livello al quale le stime devono essere prodotte, ossia i domini di stima
- la precisione che si vuole ottenere per le stime ai livello dei diversi domini
- le informazioni ausiliarie utili per la progettazione del disegno.

La *popolazione di interesse* deve essere definita sulla base di criteri che identificano precisamente le unità di analisi da sottoporre a indagine. Esempi di popolazioni sono: l'insieme delle imprese attive in Italia con riferimento a un certo periodo di tempo e appartenenti a una prefissato campo di osservazione, la popolazione delle famiglie residenti in Italia in un prefissato istante di tempo, i nati in Italia in un determinato anno solare.

L'*archivio di selezione* è una lista delle unità appartenenti alla popolazione, contenente almeno le informazioni necessarie per l'identificazione e il reperimento delle stesse. Può contenere inoltre informazioni ausiliarie utili alla fase di progettazione del disegno. In alcuni casi, l'archivio identifica gruppi di unità, o *cluster*, come ad esempio l'archivio anagrafico delle famiglie in cui la famiglia è un cluster di individui, oppure l'archivio dei comuni italiani in cui il comune è un cluster di famiglie.

Le *variabili da rilevare* possono essere di tipo qualitativo (risposte a quesiti con modalità qualitative, quali la condizione occupazionale o la percezione su un certo fenomeno) o quantitativo (quali ad esempio il reddito, la produzione o il fatturato). Di conseguenza i *parametri* oggetto di stima possono essere, come nel primo caso, frequenze assolute o relative delle modalità di risposta, oppure, come nel secondo caso, medie o totali. Il software considera in ogni caso come parametri di stima i totali di tali variabili, quindi le frequenze assolute per le variabili qualitative e i totali per le variabili quantitative.

I *domini di stima* sono le sotto-popolazioni a livello delle quali si vuole ottenere le stime dei parametri di interesse. Tali domini devono essere definiti sulla base di variabili presenti nell'archivio di selezione, ossia sulla base di variabili note a priori sulle unità della popolazione, anche incrociate tra loro. Esempi di domini di stima sono: la regione, la provincia, la regione incrociata con l'attività economica delle imprese, la classe di

età. Inoltre, è spesso necessario che le stime siano prodotte contemporaneamente per più tipi di dominio, ovvero per partizioni alternative della stessa popolazione.

La *precisione* richiesta per le stime di interesse esprime il grado di affidabilità che le stime, da produrre sulla base del campione, devono garantire. Essa viene espressa in termini di *coefficiente di variazione* (ottenuto come rapporto tra l'errore standard della stima e la stima stessa), da specificare per ciascun parametro e ciascun tipo di dominio di stima. Ad esempio, si può richiedere che la stima del totale del fatturato delle imprese a livello degli incroci di regione e classe di attività economica presenti un coefficiente di variazione non superiore al 10%. Si fa presente che, per una certa variabile, il coefficiente di variazione è lo stesso che si consideri la stima della media o del totale; per le modalità di variabili qualitative è lo stesso per la stima di una frequenza relativa e di frequenza assoluta.

Le *informazioni ausiliarie* utili per la progettazione del disegno, sono generalmente contenute nell'archivio di selezione o possono essere desunte da indagini precedenti analoghe a quella in oggetto o da un censimento. Le variabili ausiliarie necessarie per l'allocazione sono: variabili di stratificazione, indispensabili per la definizione degli strati e dei domini di stima, variabili correlate con quelle di interesse, utili per lo studio della variabilità delle stime dei parametri di interesse.

### 1.1.2. Disegno di campionamento ad uno stadio stratificato e allocazione del campione

Il software MAUSS consente di calcolare la dimensione del campione e la sua allocazione negli strati per un disegno campionario a uno stadio stratificato. Per realizzare tale schema campionario, la popolazione deve essere preliminarmente suddivisa in strati, secondo le modalità di una o più variabili di classificazione note a priori su tutte le unità presenti nell'archivio di selezione.

In una stratificazione standard gli strati costituiscono la minima partizione della popolazione che consente di ottenere i domini di stima come unione di strati (domini pianificati). In genere, stratificazioni più fini comportano un aumento della numerosità campionaria a parità di errore atteso; ciò è dovuto alla necessità di garantire almeno una o due unità campione per strato.

Per illustrare la procedura standard di costruzione degli strati si consideri, ad esempio, il caso di un'indagine sulle imprese che debba produrre le stime separatamente per classi di attività economica (individuate dalle prime quattro cifre della classificazione delle attività economiche) e classi di addetti. In tale situazione, gli strati sono definiti dalle modalità incrociate delle variabili classe di attività economica e classe di addetti.

L'allocazione del campione negli strati, i cui aspetti metodologici sono descritti in dettaglio nel paragrafo 5, viene realizzata mediante un criterio che costituisce una generalizzazione del metodo di Neyman (nota come metodo di allocazione ottima univariata) e che consente di minimizzare la dimensione campionaria avendo prefissato gli errori di campionamento attesi delle stime di interesse, relativamente a ciascun tipo di dominio di stima, realizzando un'allocazione *multi-variata* e *multi-dominio*.

Si fa presente che alcuni strati possono essere censiti in base ad una decisione del responsabile dell'indagine, (ad esempio si può decidere a priori di censire tutte le imprese con più di 20 addetti).

## *1.2. Predisposizione dei file di input*

Il software MAUSS richiede che l'utente fornisca come input le caratteristiche della popolazione oggetto di indagine, delle variabili oggetto di stima e i vincoli sull'errore campionario delle stime.

In output sono prodotte le numerosità campionarie per strato, gli errori attesi di tutte le stime di interesse e alcune informazioni utili a valutare la soluzione individuata.

Le informazioni di input devono essere fornite al software in due file di dati:

- il primo deve contenere la stratificazione della popolazione, con indicazione del numero di unità appartenenti a ogni strato, l'indicazione dei domini di stima e delle statistiche sull'intensità e sulla variabilità dei fenomeni di interesse;
- il secondo deve contenere i vincoli sugli errori campionari, specificati per ciascuna variabile di interesse e ciascun tipo di dominio di stima.

### 1.2.1. Il file degli strati

Il primo file deve contenere un record per ogni strato con le seguenti variabili (per l'indicazione dei nomi e dei formati vedi il capitolo 2 di questo manuale):

- codice di strato, $h$ ($h$=1,.., $H$);
- numero di unità della popolazione appartenenti allo strato $h$;
- codice di dominio di tipo 1, codice di dominio di tipo 2,…., codice di dominio del tipo $D$ al quale lo strato $h$ appartiene;
- medie di popolazione, calcolate a livello di strato, di ciascuna delle P variabili che si intende utilizzare per l'allocazione

$$m_{p,h} = \frac{1}{N_h} \sum_{j=1}^{N_h} Y_{p,hj} \qquad (1a)$$

in cui $Y_{pj}$ indica il valore della variabile $y_p$ sull'unità $j$ della popolazione, $N_h$ il numero di unità della popolazione appartenenti allo strato $h$; per le variabili qualitative, è necessario definire una variabile dicotomica per ogni modalità di risposta della variabile e la media corrisponde il tal caso alla frequenza relativa $f_{p,h}$ della modalità 1 della variabile dicotomica $y_p$

$$m_{p,h} = \frac{F_{p,h}}{N_h} = f_{p,h} \qquad (1b)$$

avendo indicato con $F_{p,h}$ la frequenza assoluta della medesima modalità;

- scostamenti quadratici medi di popolazione delle P variabili, calcolati a livello di strato

$$s_{p,h} = \sqrt{\frac{1}{(N_h - 1)} \sum_{j=1}^{N_h} \left(Y_{p,hj} - m_{p,h}\right)^2} \quad ; \qquad (2a)$$

per le variabili qualitative lo scostamento quadratico medio sarà calcolato come

$$s_{p,h} = \sqrt{f_{p,h}\left(1 - f_{p.h}\right)} \qquad (2b)$$

- indicazione di strato da censire (0 da campionare, 1 da censire);
- costo della rilevazione nello strato.

Per la costruzione del primo data set la principale difficoltà può derivare dal reperimento delle informazioni ausiliarie sulle variabili di interesse. Sono possibili differenti situazioni:

- medie e scostamenti quadratici medi sono desumibili dall'archivio di selezione, ma sono riferite a un periodo precedente e/o a variabili proxy delle variabili di indagine;
- medie e scostamenti quadratici medi sono ottenibili come stime da un'indagine campionaria analoga riferita a un anno precedente;
- medie e scostamenti quadratici medi sono ignoti.

Nel primo caso, che si verifica ad esempio nella situazione di un'indagine sulle imprese, per le quali può essere disponibile un archivio esaustivo di imprese (quale l'archivio ASIA, Archivio Statistico delle Imprese Attive) contenente il fatturato o il numero di addetti per ciascuna impresa con riferimento a un anno precedente, è immediato calcolare a livello di strato le quantità richieste, secondo le espressioni (1a), (1b), (2a) e (2b). Spesso le variabili disponibili sull'archivio costituiscono una proxy delle variabili oggetto di indagine e se la correlazione tra le variabili ausiliarie e le variabili di interesse è sufficientemente elevata, si può garantire un buon livello di precisione sulle stime delle variabili di interesse (Cicchitelli et al.,1992).

Nel secondo caso è possibile ottenere la stima delle medie e scostamenti quadratici medi di popolazione dai dati campionari di un'indagine effettuata in un periodo precedente. In questo caso è anche opportuno valutare il grado di affidabilità di tali stime e utilizzarle a un livello di aggregazione superiore rispetto allo strato qualora esse non presentino una precisione sufficiente a livello di strato. Di tale situazione ne sarà illustrato un esempio nel seguito.

La terza situazione sussiste quando non si dispone di alcuna informazione sulla variabilità dei fenomeni di interesse perché l'indagine in corso di progettazione viene condotta per la prima volta. In questi casi si può impostare la procedura di allocazione stabilendo quali sono i domini di stima e considerando un numero di stime di frequenze relative "tipiche", sufficienti a coprire il campo di variazione di tutte le stime che l'indagine ha l'obiettivo di produrre. Ad esempio se un'indagine ha l'obiettivo di produrre stime a livello di regione, ripartizione e totale nazionale, si può stabilire che il disegno campionario debba essere tale da garantire una sufficiente precisione per stime corrispondenti almeno all'1% a livello nazionale, al 3% a livello di ripartizione geografica e al 5% a livello regionale. In tal caso gli strati saranno costituiti dal dominio più disaggregato, ossia la regione, e si utilizzano tre variabili le cui medie saranno costanti per tutti gli strati:

$f_{1,h} = 0.01,\ f_{2,h} = 0.03,\ f_{3,h} = 0.05$ per ogni strato *h,*

mentre gli scostamenti quadratici medi si otterranno in base alla formula (2b).

Il software MAUSS fornirà la dimensione campionaria complessiva e la sua allocazione tra le regioni in modo tale da rispettare i vincoli sull'errore campionario delle stime "tipiche" prefissate a livello dei differenti domini.

## 1.2.2. Il file dei vincoli sugli errori campionari

Il secondo file deve contenere un record per ogni tipo di dominio con le seguenti variabili:

- codice di tipo di dominio, d (d=1,.., D);
- valore massimo ammesso del coefficiente di variazione per la stima del totale delle K variabili di interesse, cv1,…,cvK.

La predisposizione del secondo file richiede che l'utente specifichi per ognuna delle stime di interesse il valore massimo del coefficiente di variazione ammesso a livello di ognuna delle tipologie di domini di stima definiti.

E' utile precisare che se per una certa stima non è richiesto che sia rispettato un limite per l'errore campionario a livello di un certo tipo di dominio, si può indicare un valore molto elevato del coefficiente di variazione per quel tipo di dominio, come ad esempio cv=1.

Riguardo al criterio con cui viene fissato il livello dell'errore per dominio, è pratica comune allocare il campione in modo che tale valore sia approssimativamente uguale per tutti i domini (Sigman e Monsour, 1995). In alternativa, è stato proposto (Bankier, 1988; Hidiroglou et al., 1995) un metodo di allocazione più generale che si basa sulla determinazione del valore del coefficiente di variazione per dominio in maniera iterativa, ipotizzando diversi livelli di aggregazione degli strati. In particolare, per i livelli più bassi, tale coefficiente viene fissato in maniera inversamente proporzionale alla stima del totale della variabile d'interesse. Per i successivi livelli di aggregazione, che corrispondono ai domini, il coefficiente di variazione è ottenuto, a partire dai valori precedentemente calcolati, attraverso la risoluzione di un algoritmo iterativo che tiene contemporaneamente conto dei vincoli (determinati per i livelli più bassi) e del fatto che ciascun dominio deve avere approssimativamente lo stesso coefficiente di variazione.

## 1.2.3. Esempio di costruzione dei data set di input

A titolo di esempio si illustra la stratificazione utilizzata per l'indagine ISTAT sulle nascite. La popolazione obiettivo è costituita dalle madri dei nati in un certo anno solare, stratificate per cinque classi di età e regione.

I domini di stima sono la regione, la ripartizione geografica, la classe di età a cinque modalità e il dominio nazionale. Supponiamo per semplicità che le stime di interesse siano due: la frequenza relativa delle donne che erano occupate prima del parto ma non più occupate al momento dell'intervista e la frequenza relativa delle donne i cui figli frequentano l'asilo nido. L'informazione sulle variabili di interesse è in questo caso desumibile dai dati di un'indagine precedente, ma viene utilizzata non a livello di singolo strato, in quanto ritenuta non sufficientemente affidabile, bensì a livello di ripartizione geografica e classi di età. La variabile relativa al costo di ciascuno strato è posta pari a uno in quanto non esiste una differenziazione di costo tra i diversi strati. Lo stesso per la variabile che indica la presenza di strati da censire: in questa indagine presenta sempre valore zero in quanto non si ritiene utile censire alcuno strato.

Il file degli strati ha la struttura riportata nella tabella seguente.

| STRATO | Dom1 = Regione | Dom 2 = Classe di età | Dom 3 = Rip | Dom 4 = Totale | Pop | Media 1 | SQM1 | Media 2 | SQM2 | Costo | Cens |
|---|---|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 15-24 Piemonte | Piemonte | 15-24 | Nord Ovest | 1 | 1000 | 0.20 | 0.4 | 0.45 | 0.497 | 1 | 0 |
| 25-29 Piemonte | Piemonte | 25-29 | Nord Ovest | 1 | 1600 | 0.18 | 0.348 | 0.5 | 0.5 | 1 | 0 |
| … | | | | | | | | | | | |
| … | | | | | | | | | | | |
| 35-39 Sardegna | Sardegna | 35-39 | Isole | 1 | 700 | 0.30 | 0,458 | 0.2 | 0.4 | 1 | 0 |
| 40 e oltre Sardegna | Sardegna | 40 e oltre | Isole | 1 | 300 | 0.30 | 0,458 | 0.25 | 0.433 | 1 | 0 |

Il secondo file, contenente i vincoli sugli errori campionari, presenta la seguente struttura.

| **Tipo di dominio** | **CV1** | **CV2** |
|---|---|---|
| Dom1 = Regione | 0.10 | 0.14 |
| Dom 2 = Classe di età | 0.06 | 1 |
| Dom 3 = Rip | 0.05 | 0.08 |
| Dom 4 = Totale | 0.02 | 0.03 |

I valori attribuiti ai coefficienti di variazione massimi per le due stime a livello dei quattro tipi di domini sono puramente esemplificativi, ma mostrano come, in generale, alle tipologie di domini con un numero maggiore di modalità, venga attribuito un valore più elevato del coefficiente di variazione. Si fa notare che poiché per la seconda variabile non è richiesta la stima a livello delle classi di età (dom2), il vincolo è stato posto pari ad 1.

## *1.3. L'utilizzo dell'output della procedura*

Il sistema produce come output: (a) le numerosità campionarie per strato, (b) gli errori attesi di campionamento per ciascun incrocio variabile-dominio d'interesse; (c) alcune statistiche utili per la messa a punto del piano di campionamento.

Le numerosità campionarie per singolo strato sono aggiunte al data set degli strati, mentre gli errori attesi di campionamento sono riportati sia nel data set di uscita sia nelle tabelle 7 e 8 stampate in output. Le statistiche per la messa a punto sono invece riportate nella tabella 5.

Il sistema prevede che la soluzione finale possa essere scelta confrontando i risultati relativi a più prove, ottenute definendo in modo alternativo l'input richiesto in termini di errori campionari. In questo modo il software si propone anche come uno strumento di ausilio alla progettazione stessa dell'indagine. Infatti, effettuare simulazioni al variare degli input consente ai responsabili delle indagini di definire meglio sia gli obiettivi dell'indagine sia le risorse necessarie per conseguirli.

Lo strumento a disposizione dell'utente per modificare i dati di input, in particolare i dati contenuti nel secondo file di input, è la tabella 5, contenente l'analisi di sensitività. In tale tabella per ogni stima e ogni tipo di dominio è riportato il valore della numerosità

campionaria aggiuntiva necessaria per conseguire un decremento del 10% del coefficiente di variazione della corrispondente stima. Tale numero può essere interpretato anche in senso opposto, cioè come la diminuzione di numerosità campionaria che si otterrebbe incrementando del 10% l'errore della corrispondente stima a livello di quel tipo di dominio.

Ad esempio, riprendendo l'esempio relativo all'indagine sulle nascite, supponiamo che la sensitività della stima della variabile 1 a livello del primo tipo di dominio (regione) sia di 567 unità. Poiché il coefficiente di variazione di tale stima era stato impostato al 10% (cv1=0.10), ciò significa che

- per ottenere una diminuzione della numerosità campionaria di 567 unità è necessario portare il cv1 da 0.10 a 0.11, ossia aumentare l'errore campionario atteso del 10%,
- che per diminuire l'errore campionario atteso del 10%, è necessario aggiungere al campione 567 unità.

Utilizzando tale strumento l'utente è messo in condizioni di effettuare gli aggiustamenti necessari per raggiungere la numerosità campionaria desiderata o, viceversa, di raggiungere i vincoli sugli errori campionari desiderati.

## *1.4. La metodologia di allocazione multivariata e multi-dominio*

In generale, con la specificazione della numerosità campionaria negli strati ci si propone di minimizzare la variabilità delle stime. In assenza di informazioni specifiche sulla variabilità negli strati, l'obiettivo viene raggiunto attraverso l'allocazione proporzionale; avendo la possibilità di ricorrere, invece, ad informazioni ausiliarie è possibile definire allocazioni più efficienti.

Nel caso di un'unica variabile di interesse, disponendo di una stima della variabilità, si può far riferimento ai risultati ben noti per l'allocazione ottima nel caso univariato (Cochran, 1977); questi risultati permettono di determinare la numerosità in modo da minimizzare la varianza di stima per un fissato valore della funzione di costo o, viceversa, da minimizzare i costi, avendo fissato precedentemente il livello di accuratezza delle stime.

La soluzione univariata non è comunque idonea per la progettazione della maggior parte delle indagini sulle imprese, nelle quali solitamente si rilevano più variabili d'interesse. Per queste indagini, pertanto, è necessario affrontare il problema dell'allocazione ottima nell'ambito di un approccio multivariato. La seguente trattazione è ripresa da Falorsi *et al.* (1998).

### 1.4.1. Problema dell'allocazione multivariata

In un campione stratificato con selezione delle unità con probabilità uguali e senza reimmissione, la varianza dello stimatore del totale della generica variabile d'interesse $y_p$ (p=1,...,P) può essere espressa come:

$$V_p' = V_p + V_{0p} = \sum_{h=1}^{H} \frac{N_h^2}{n_h} S_{p,h}^2 - \sum_{h=1}^{H} N_h S_{p,h}^2 \qquad (3)$$

dove $S^2_{p,h}$ è la varianza della variabile p nello strato h e $V_{0p}$ denota la parte di varianza non influenzata dall'allocazione.

Si definisce, inoltre, la seguente funzione di costo:

$$C' = C_0 + C = C_0 + \sum_{h=1}^{H} C_h n_h \qquad (4)$$

dove $C_0$ rappresenta il costo fisso dell'indagine che non dipende né dalla numerosità campionaria né dall'allocazione, C il costo variabile e $C_h$ (h=1,...,H) il costo per unità campionaria relativo allo strato h.

E' possibile determinare la numerosità da assegnare a ciascuno strato secondo due approcci (Sigman e Monsour, 1995). Il primo consiste nel minimizzare il prodotto W $\times$C, dove $W = \sum_{p=1}^{P} W_p V_p$, indicando con $W_p$ (p=1,...,P) dei pesi da definire. La soluzione viene trovata fissando il valore di W o di C. Questo metodo risulta poco applicabile nelle situazioni concrete per la difficoltà di specificare in maniera non arbitraria i pesi.

Nel secondo approccio si fissa un estremo superiore, $V_p^*$, per ciascun $V_p'$ e si minimizza la funzione di costo C sotto i vincoli $V_p' \leq V_p^*$ (p=1,..,P).
Il software MAUSS utilizza questo secondo approccio, adottando una generalizzazione della soluzione proposta da Bethel (1989), in cui viene definito un problema di minimo vincolato con funzione obiettivo convessa e vincoli di tipo lineare. In particolare, viene riformulata la quantità C della (4) ponendo:

$$x_h = \begin{cases} 1/n_h & \text{se } n_h \geq 1 \\ \infty, & \text{altrimenti} \end{cases};$$

In questo modo l'espressione della funzione obiettivo da minimizzare diviene:

$$f(\mathbf{x}) = \sum_{h=1}^{H} C_h / x_h \qquad (5)$$

dove $\mathbf{x} = (x_1,..., x_H)'$. I vincoli $V_p' \leq V_p^*$ assumono la forma:

$$\sum_{h=1}^{H} a_{p,h} x_h \leq 1 \quad , \; p=1,...,P \qquad (6)$$

essendo

$$a_{p,h} = \frac{N_h^2 S_{p,h}^2}{\left(V_p^* - V_{0p}\right)} \quad . \tag{7}$$

Dal momento che il problema di minimizzazione della (5) sotto i vincoli (6) soddisfa le condizioni del teorema di Kokan e Khan (1967) esiste una soluzione ottima $\mathbf{x}^*$. Utilizzando il teorema di Kuhn-Tucker (1951), Bethel dimostra che esistono dei valori $\lambda_p^* \geq 0$, tali che la soluzione ottima assume la forma:

$$x_h^* = \sqrt{C_h} \Bigg/ \left( \sqrt{\sum_{p=1}^{P} \mu_p^* a_{p,h}} \sum_{k=1}^{H} \sqrt{C_k \sum_{p=1}^{P} \mu_p^* a_{p,k}} \right) \tag{8}$$

dove $\mu_p^* = \lambda_p^* \Big/ \sum_{p=1}^{P} \lambda_p^*$ per cui $\sum_{p=1}^{P} \mu_p^* = 1$. (9)

Per determinare simultaneamente i valori ottimi $x_h^*$ e $\mu_p^*$ è necessario ricorrere ad algoritmi di risoluzione numerica, come quelli proposti nel lavoro di Bethel .

### 1.4.2. Allocazione multivariata per più domini e per più tipi di dominio

La soluzione illustrata nel precedente paragrafo è relativa al caso in cui le stime dei parametri d'interesse debbano essere fornite a livello dell'intera popolazione; in generale, però, le indagini campionarie hanno l'obiettivo di fornire le stime non solo per l'intera popolazione, ma anche per sottopopolazioni (*domini di studio*) individuate da una partizione (o *tipo di dominio*) della popolazione oggetto d'indagine. Inoltre, è spesso necessario che le stime siano prodotte contemporaneamente per più tipi di dominio, ovvero per partizioni alternative della stessa popolazione. In questi casi, il campione deve essere pianificato in modo tale da assicurare *simultaneamente* l'accuratezza delle stime ai diversi livelli di dettaglio richiesti e ciò può essere ottenuto generalizzando la soluzione descritta in precedenza.

Per illustrare il metodo utilizzato per l'allocazione multivariata nel caso di più domini di stima, si indichi con: d (d=1,...,D), il generico tipo di dominio; $k_d$ ($k_d$=1,...,$K_d$), il generico dominio di tipo d; $H_{k_d}$, il numero di strati che appartengono al dominio $k_d$. La funzione obiettivo (5) rimane invariata, mentre il sistema dei vincoli può essere ridefinito nel modo seguente:

$$\sum_{h=1}^{H_{k_d}} \frac{N_h^2}{n_h} S_{p,h}^2 - \sum_{h=1}^{H_{k_d}} N_h S_{p,h}^2 \leq V_{p,k_d}^* \quad (p=1,...,P;\ d=1,...,D;\ k_d=1,...,K_d) \tag{10}$$

dove $V_{p,k_d}^*$ è il limite superiore della varianza della stima del totale della variabile p per il dominio $k_d$.

Analogamente a quanto fatto nel paragrafo 5.1, la (10) può essere scritta come:

$$\sum_{h=1}^{H} a_{p,k_d,h}\, x_h \le 1 \qquad (p=1,...,P;\ d=1,...,D;\ k_d=1,...,K_d)$$

dove

$$a_{p,k_d,h} = \frac{N_h^2 S_{p,h}^2 \delta_{k_d,h}}{\sum_{h=1}^{H} N_h S_{p,h}^2 \delta_{k_d,h} + V_{p,k_d}^*}, \tag{11}$$

$$\text{con } \delta_{k_d,h} = \begin{cases} 1 & \text{se } h \in k_d \\ 0 & \text{altrimenti} \end{cases}.$$

Se si definisce un indice r i cui valori sono in corrispondenza biunivoca con i valori individuati dall'ordinamento lessicografico del vettore identificato dai tre indici (d, $k_d$, p), il sistema dei vincoli diviene:

$$\sum_{h=1}^{H} a_{r,h} x_h \le 1 \qquad \text{per } r=1,...,R, \text{ dove } R = P\sum_{d=1}^{D} K_d\ , \tag{12}$$

ovvero una forma del tutto equivalente alla (6).

Riprendendo la (8), ed essendo ancora soddisfatte le condizioni dei teoremi di Kokan-Khan e Kuhn-Tucker, la soluzione ottima che minimizza la (5) sotto i vincoli (12) è:

$$x_h^* = \sqrt{C_h} \Bigg/ \left( \sqrt{\sum_{r=1}^{R} \mu_r^* a_{r,h}} \sum_{k=1}^{H} \sqrt{C_k \sum_{r=1}^{R} \mu_r^* a_{r,k}} \right) \tag{13}$$

$$\text{dove } \mu_r^* = \lambda_r^* \Bigg/ \sum_{r=1}^{R} \lambda_r^* \quad \text{con} \quad \sum_{r=1}^{R} \mu_r^* = 1. \tag{14}$$

## 1.4.3. Algoritmi di risoluzione

L'algoritmo proposto da Bethel per il calcolo dell'allocazione ottima nel caso multivariato può essere generalizzato per la soluzione del medesimo problema in presenza di più tipi di dominio. Questo algoritmo trova la soluzione ottima in maniera iterativa, partendo da una soluzione iniziale ($\nu$=1) che coincide con la soluzione ottima nel caso univariato per la prima variabile sul primo dominio (r=1). In genere con questa soluzione la funzione obiettivo assume un valore molto piccolo e i rimanenti vincoli (r=2,..,R) non sono soddisfatti. In ciascuno dei passi successivi, $\nu$=2,3,... , la numerosità campionaria viene aumentata, incrementando la funzione obiettivo $f\left(\mathbf{x}^{(\nu)}\right) \ge f\left(\mathbf{x}^{(\nu-1)}\right)$, fino al soddisfacimento di tutti i vincoli. Bethel dimostra che tale algoritmo converge e che, quindi, è possibile individuare simultaneamente $\nu^*$ e $x^*$ in modo che $0 \le f\left(\mathbf{x}^{(\nu)}\right) \le f(\mathbf{x}^*)$.

La complessità computazionale di tale algoritmo, in particolare nel caso di più domini di stima, ha determinato il ricorso all'algoritmo proposto da Chromy (1987) che risulta di più immediata implementazione e sembra convergere verso la soluzione ottima in modo più veloce. Per illustrare tale algoritmo si indichi con: $\mathbf{A}=\left\{a_{r,h}\right\}$ la matrice di dimensioni R e H i cui elementi sono definiti dalla (11)[1] e con $\mathbf{a}_r$ l'r-esima riga di **A**. L'algoritmo di Chromy è un algoritmo iterativo che al primo passo calcola il valore di **x** in base alla (13), ponendo ogni elemento di $\boldsymbol{\mu}$ pari a 1/R. Se tale soluzione soddisfa tutti i vincoli, l'algoritmo si arresta. In caso contrario, l'algoritmo calcola $\mathbf{x}^{(\nu)}$ in corrispondenza del vettore $\boldsymbol{\mu}^{(\nu)}$, il cui elemento generico è fornito dalla seguente espressione:

$$\mu_r^{(\nu)} = \mu_r^{(\nu-1)}\left(\mathbf{a}_r\mathbf{x}\left(\boldsymbol{\mu}^{(\nu-1)}\right)\right)^2 \Big/ \sum_{r=1}^{R} \mu_r^{(\nu-1)}\left(\mathbf{a}_r\mathbf{x}\left(\boldsymbol{\mu}^{(\nu-1)}\right)\right)^2 \qquad 1 \le r \le R \qquad (15)$$

dove $\mathbf{x}\left(\boldsymbol{\mu}^{(\nu-1)}\right)$ denota il valore di **x**, ottenuto sulla base della (13) ponendo $\boldsymbol{\mu} = \boldsymbol{\mu}^{(\nu-1)}$.

Poiché questi algoritmi non assicurano che la soluzione ottima sia tale che $n_h \le N_h$, MAUSS contiene una procedura di riallocazione iterativa che prevede il censimento degli strati in cui $n_h > N_h$ e calcola nuovamente le numerosità campionarie sotto le mutate condizioni.

[1] Le dimensioni di questa matrice sono analoghe a quelle proposte dall'approccio di Causey (1983).

# 2. **MAUSS: utilizzo dello strumento**

## *2.1 Installazione*

Ambiente Microsoft Windows.

I requisiti hardware minimi per Mauss-R sono:

RAM: almeno 512MB

Spazio disco: 5MB

Inoltre è necessario che sul PC siano installati:

Java 2 Runtime Environment versione 6.0 o superiore scaricabile da http://java.sun.com/javase/downloads/index.jsp

Ambiente R versione 7.0 o superiore scaricabile da http://cran.r-project.org/bin/windows/base/

La variabile di ambiente PATH deve contenere i puntamenti agli eseguibili java.exe e R.exe.

Per modificare la variabile PATH:

*Start -> Impostazione -> Pannello di controllo -> Sistema -> Avanzate -> Variabili d'ambiente*

A questo punto selezionare la variabile PATH e cliccare sul bottone *modifica.* Qui aggiungere, all'inizio della stringa, il percorso della cartella che contiene il file *java.exe* e quello della cartella che contiene *R.exe* separati da *";"*.

Per esempio:

PATH=*C:\Programmi\Java\jre1.6.0_03\bin;C:\Programmi\R\R-2.7.1\bin;*
C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;

**Installazione**

Per installare il software si deve scaricare il file *setup_MaussR.exe* sul proprio PC ed eseguirlo.

## *2.2 Utilizzo del software*

### 2.2.1. Avvio della procedura

Da Menu:

Start ->Programmi->mauss->MaussR

Da Desktop: doppio-click sull'icona 

### 2.2.2. Menu principale

Dopo l'avvio della procedura, appare la schermata principale con il menu con le seguenti funzioni:


**Figura 1 - Menu principale**

**File** - Definizione del progetto: creazione di un nuovo progetto, apertura di un progetto esistente, chiusura del progetto in corso e uscita dall'applicazione.

**Definizione** - Definizione dei parametri per l'elaborazione e gestione dei vincoli.

**Allocazione** - Calcolo dell'allocazione ottimale con il metodo di Bethel per la versione corrente del file dei vincoli o per tutte le versioni.

**Report** - Visualizzazione dei risultati e delle stampe.

**Help** - Visualizzazione dell'help on-line.

### 2.2.3. Menu File - Definizione del progetto

Un progetto, per MAUSS-R, è individuato dalla cartella di lavoro, cioè dalla cartella che conterrà tutti i file generati dall'applicazione. Altre informazioni necessarie all'individuazione del progetto sono i nomi dei file di input che devono essere preparati dall'utente: il file contenente le informazioni sugli strati e sulle variabili di interesse (medie, scarto quadratico medio) e quello con i vincoli desiderati (coefficienti di variazione per le stime). Per la descrizione dei due file vedi sotto "Dati input del software".

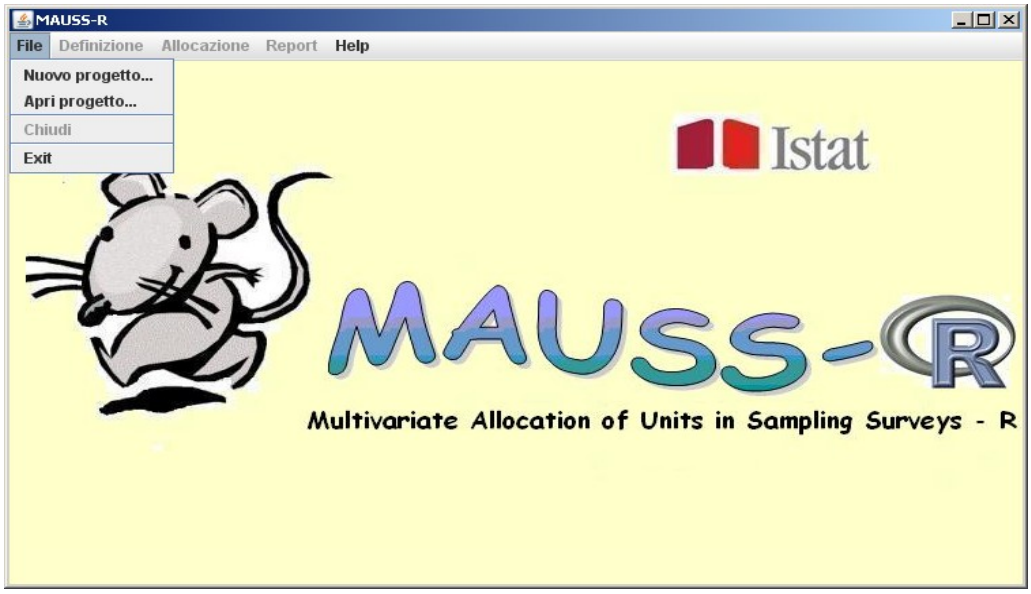



**Figura 2 - Menu File**

Funzioni:

**Nuovo progetto**: Inserimento di un nuovo progetto.

Appare la maschera in cui si richiedono la cartella di lavoro e i nomi dei due file di input.

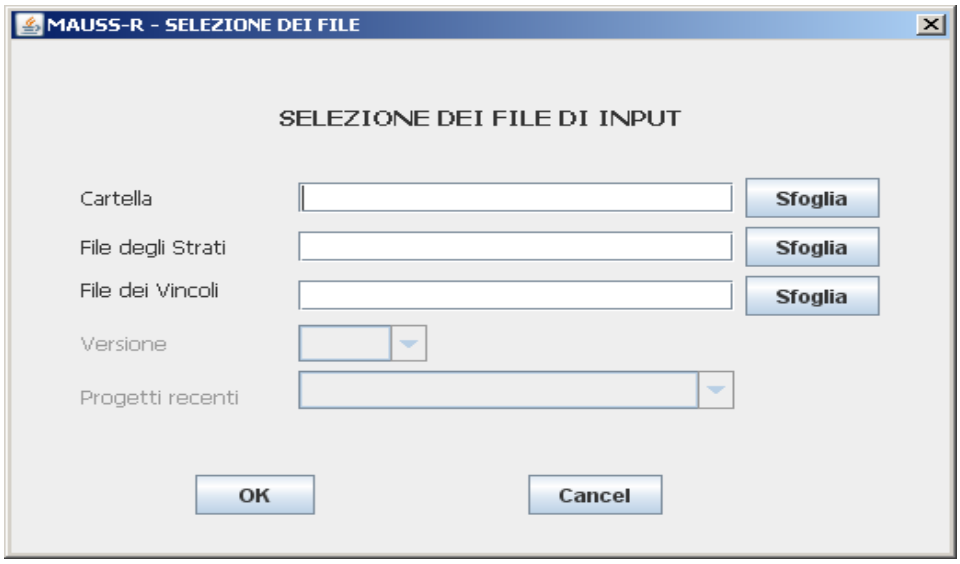



**Figura 3 - Nuovo progetto**

I nomi dei file possono essere inseriti direttamente nella casella di testo oppure essere selezionati tramite File Manager utilizzando il bottone Sfoglia.

Dopo aver dato la conferma (OK), la procedura provvede a controllare i dati inseriti e, se tutto è a posto, a predisporre l'ambiente: imposta il progressivo della versione del file dei vincoli a 1 e crea una sottodirectory della cartella di lavoro dal nome BethV1 dove copia il file dei vincoli e dove saranno scritti i risultati dell'allocazione relativi al file dei vincoli versione 1.

Se nella cartella di lavoro scelta sia già stato definito un progetto, il sistema chiede se si vuole continuare creando un nuovo progetto. In caso affermativo, ripulisce la cartella spostando tutti i risultati della precedente elaborazione in una sotto-cartella dal nome backupNNNNNN dove NNNNNN è un numero che rappresenta l'ora di sistema espressa in millisecondi. Altrimenti chiude la maschera senza definire il progetto che potrà essere aperto usando la funzione Apri progetto.

**Apri progetto**: Apertura di un progetto esistente

In questo caso, nella maschera per la definizione del progetto sono abilitati i campi per la selezione della versione corrente del file dei vincoli e la lista dei progetti precedentemente elaborati.

Scegliendo un progetto dalla lista, i campi per la scelta della cartella di lavoro e dei due file di input sono valorizzati automaticamente e non possono essere modificati.

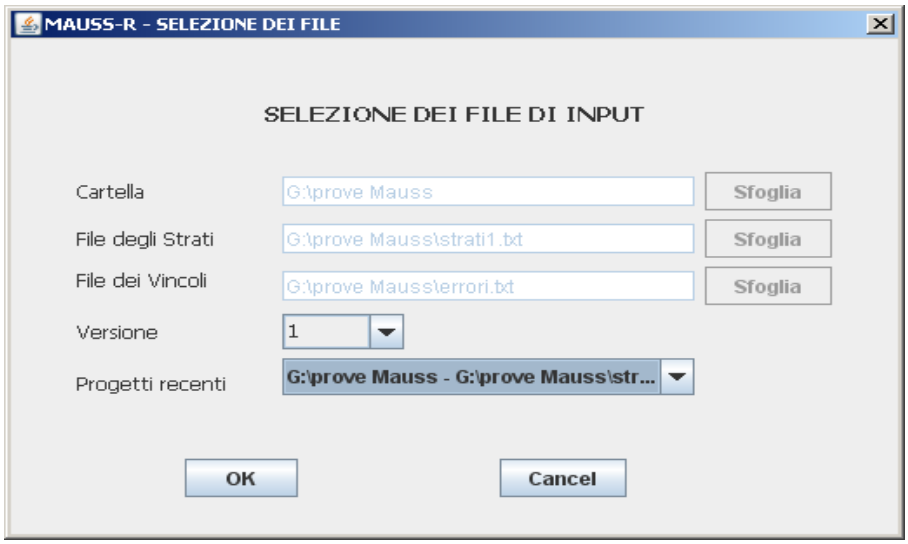



**Figura 4 - Apri progetto**

**Chiudi**: Chiusura del progetto in corso

**Exit**: Uscita dall'applicazione

### 2.2.4. Menu Definizione

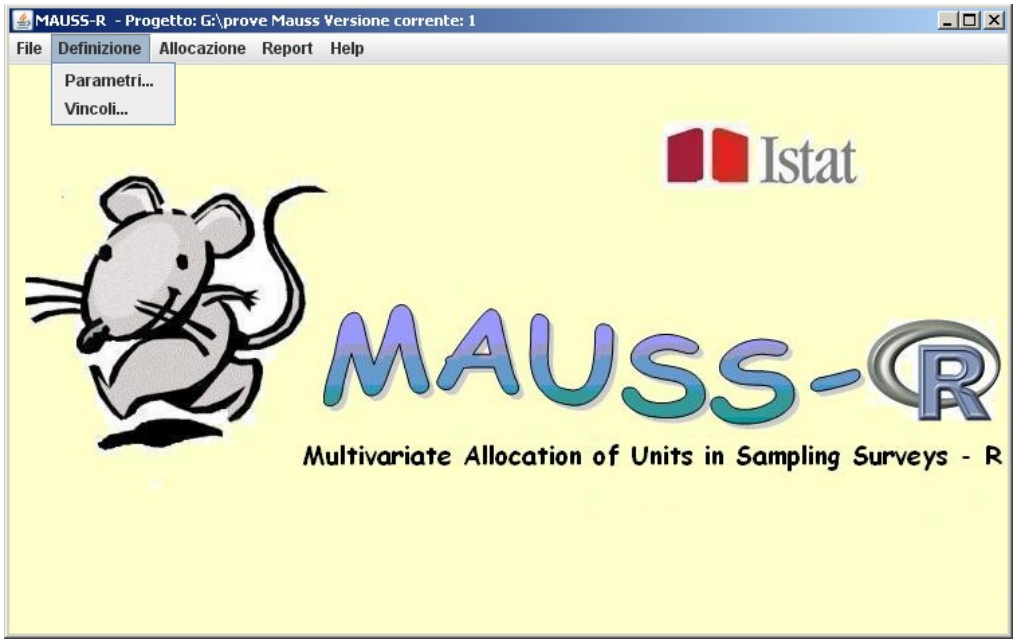



**Figura 5 - Menu Definizione**

**Funzioni:**

**Parametri:** Definizione dei parametri di elaborazione**.**

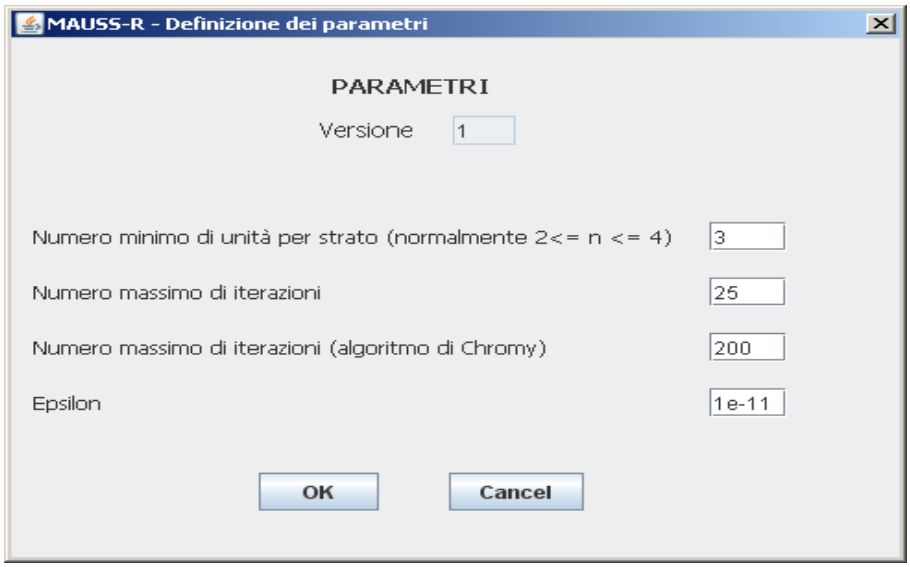



**Figura 6 - Definizione parametri**

Si possono modificare i seguenti parametri:

- Il numero minimo di unità per strato (default=2);
- Il numero massimo di iterazioni (default=25) della procedura. Queste iterazioni possono essere necessarie nel caso in cui ci siano strati in cui il numero di unità

da allocare calcolato è maggiore o uguale alla popolazione. Lo strato è impostato come censuario e la procedura viene re-inizializzata;

- Il numero Massimo di iterazione nell'algoritmo di Chromy (default=200);
- Epsilon (default=1e-11): questo valore è usato per confrontare le differenze fra i risultati nelle diverse iterazioni; se la differenza è minore di epsilon la procedura termina.

**Vincoli**: Gestione del file dei vincoli.

Permette la modifica della versione corrente dei valori dei coefficienti di variazione nonché l'inserimento di una nuova versione.

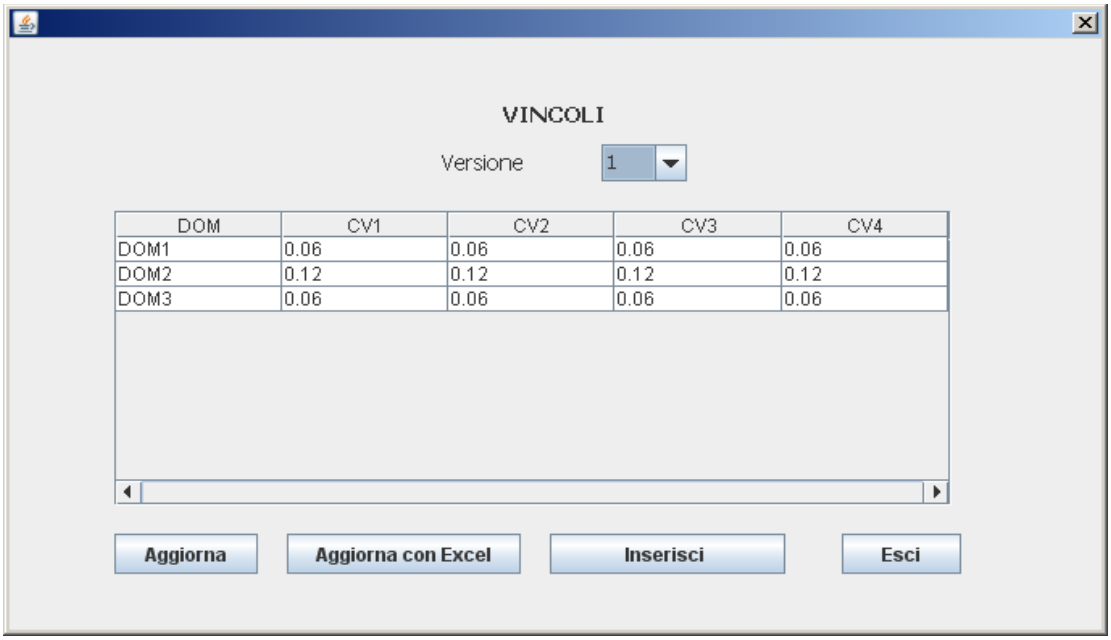



**Figura 7 - Definizione vincoli**

Per cambiare un coefficiente di variazione bisogna posizionarsi sulla casella, scrivere il nuovo valore e passare alla casella successiva con il tasto tab o con il mouse.

ATTENZIONE! La variazione non viene registrata se il cursore rimane posizionato nella casella modificata.

Il bottone Aggiorna permette la modifica della versione corrente del file.

Il bottone Inserisci aggiorna il numero delle versioni del file dei vincoli, crea una nuova sottocartella della directory di lavoro dal nome BethVn dove n è il numero della nuova versione e inserisce i dati visualizzati nella tabella in un nuovo file dei vincoli.

Per cambiare la versione corrente del file si usa il list-box Versione.

### 2.2.5. Menu Allocazione

Lancia il programma R che calcola l'allocazione campionaria nel caso multivariato per più domini di stima per le indagini ad uno stadio di campionamento.


**Figura 8 - Menu Allocazione**

Funzioni:

**Versione corrente**: Calcolo dell'allocazione ottimale per la versione corrente.

**Tutte le versioni**: Calcolo dell'allocazione ottimale per tutte le versioni del file dei vincoli.

### 2.2.6. Menu Report

Visualizzazione delle stampe relative a informazioni generali sulla popolazione e ai risultati dell'allocazione.

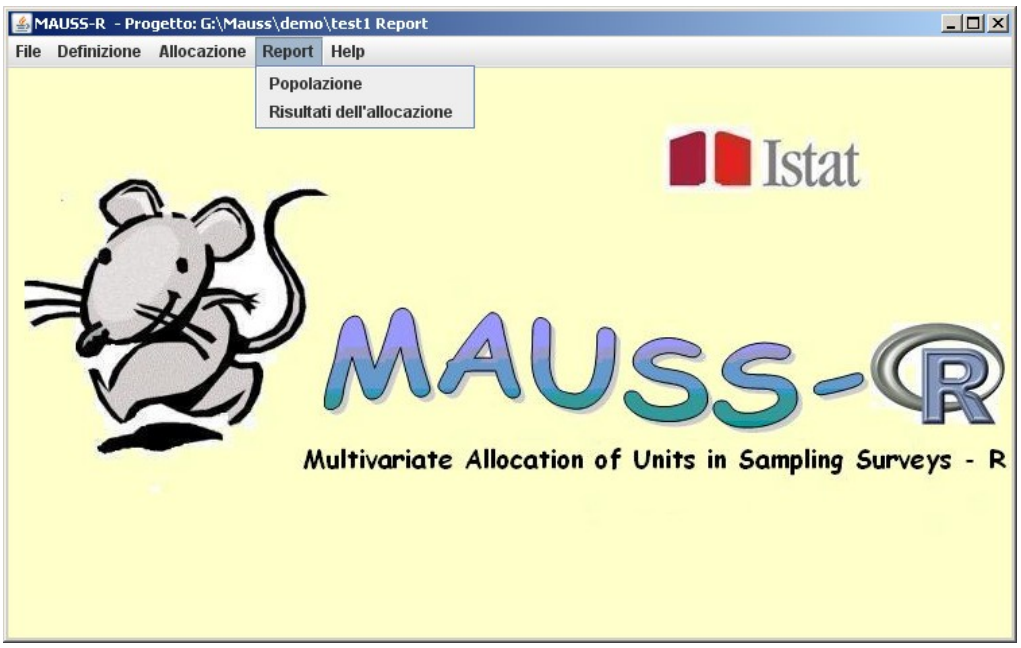



**Figura 9 - Menu Report**

Funzioni:

**Popolazione**: Analisi della popolazione.

In questa finestra (vedi fig. 10) sono visualizzate due tabelle con le informazioni sulla popolazione. La prima contiene dati di carattere generale come la numerosità della popolazione, il numero di variabili rilevate, il numero di strati, il numero dei diversi tipi di dominio. La seconda è una tabella con la popolazione e il numero degli strati per ogni dominio.

Queste tavole sono registrate nel file: *Bethel_Report_Pop1.xls*


**Figura 10 – Informazioni sulla popolazione**

### Risultati dell'allocazione

In questa finestra sono visualizzate tre tavole per l'analisi dei risultati dell'allocazione.

**Risultato dell'allocazione per la versione corrente del file dei vincoli:**

pe ogni strato, il valore della dimensione del campione ottenuta con il metodo di Bethel è confrontata con quella ottenuta con un'allocazione proporzionale e uguale negli strati.

Questa tavola è registrata nel file: *Bethel_Report1.xls*

**Confronto fra allocazioni**: Tabella di confronto fra i risultati dell'allocazione al variare dei vincoli.

In questa tabella sono riportate, per ogni strato, le numerosità campionarie ottenute con il metodo dell'allocazione ottimale di Bethel per le diverse versioni del file dei vincoli.

Nome File: *BethelResults.xls*

**Sensibilità**: Coefficienti di variazione e sensibilità.

In questa tabella sono riportati, per ogni dominio, i coefficienti di variazione attesi e effettivi e la sensibilità alla variazione del 10% della precisione desiderata.

Nome File: *Bethel_Report2.xls*

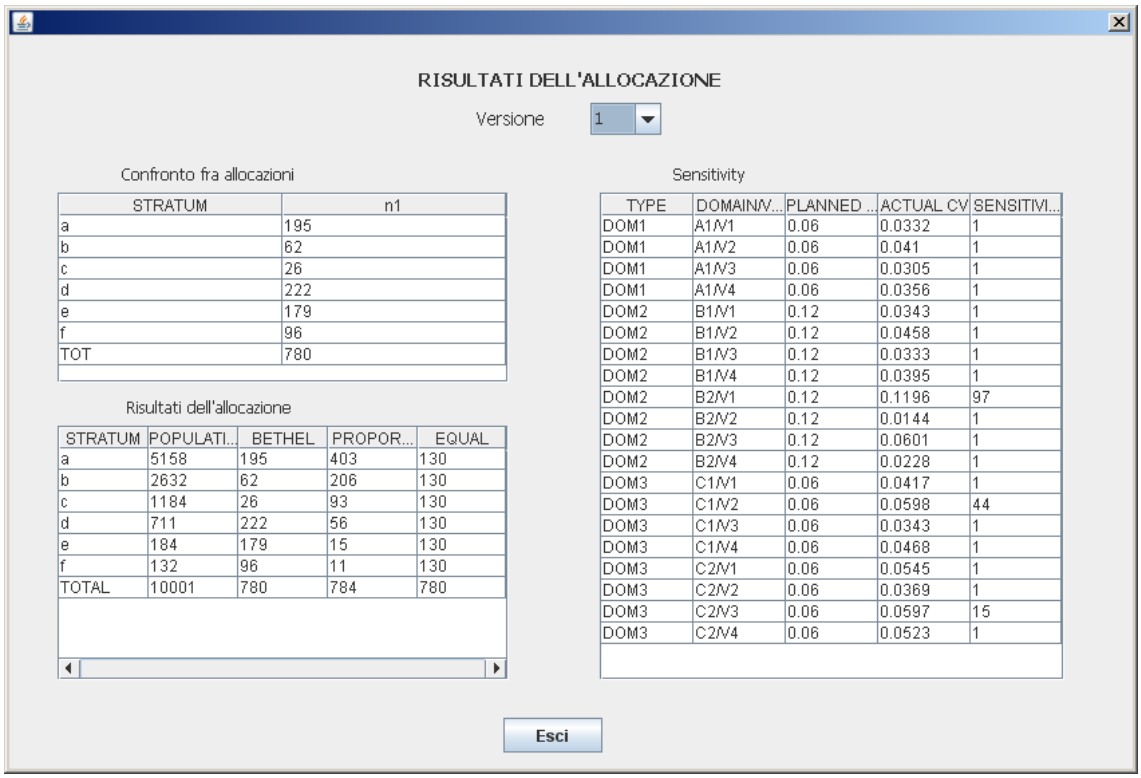



**Figure 12 – Risultati dell'allocazione**

## *2.3. Dati di input*

### 2.3.1. File degli strati

**Formato del file**: delimitato da tabulatore (.txt).

**Testata**: La prima riga del file deve contenere i nomi delle variabili specificati nella tabella in qualsiasi ordine. Il file può contenere anche altre variabili.

**Dati**: un record per ogni strato contenente le informazioni relative a tutte le variabili (le variabili COST e CENS possono essere tralasciate). Possono essere presenti anche dati relativi ad altre variabili non coinvolte nel calcolo dell'allocazione ottimale.

| **Nome della variabile** | **Descrizione** | **Formato** |
|---|---|---|
| STRATO | Codice dello strato | A |
| N | Numerosità della popolazione dello strato | N |
| DOM1, DOM2, …,DOMp | Codici dei domini (1…p) | A |
| M1, M2, …, Mn | Media per le *n* variabili nella popolazione | N |
| S1, S2, …, Sn | Deviazione standard per le *n* variabili nella popolazione | N |
| COST | Costo unitario per lo strato. Default=1 | N |
| CENS | Copertura dello strato:<br>1 = censuario.<br>0 = campionario.<br>Default = 0. | N |

### 2.3.2. File dei vincoli di precisione delle stime

**Formato del file:** delimitato da tabulatore (.txt).

**Testata:** La prima riga del file deve contenere i nomi delle variabili specificati nella tabella in qualsiasi ordine. Il file può contenere anche altre variabili.

**Dati:** Coefficienti di variazione per tutti i tipi di dominio. Un record per ogni tipo di dominio contenente le informazioni relative alle variabili elencate nella tabella. Possono essere presenti anche dati relativi ad altre variabili non coinvolte nel calcolo dell'allocazione ottimale.

| Nome della variabile | Descrizione | Formato |
|---|---|---|
| DOM | Type of domain code. | A |
| CV1, CV2, …, CVn | Planned coefficient of variation for n variables. | N |

## *2.4. Output del software*

**Formato del file:** delimitato da tabulatore (.txt).

**Nome del file:** Bethel_campio.txt

**Cartella:** Sotto-cartella della directory di lavoro relative alla versione.

E' una copia del file degli strati di input a cui viene accodata la variabile CAMP con il risultato dell'allocazione ottimale di Bethel.

## *2.5. File di supporto per l'applicazione grafica*

### 2.5.1. Elenco dei progetti

**Formato del file**: file delimitato da ";"

**Nome File**: Progetti.csv

**Cartella**: $HOME/.Mauss2

| Nome variabile | Descrizione | Tipo |
|---|---|---|
| folder | Cartella di lavoro | A |
| strati | Nome del file degli strati | A |
| vincoli | Nome del file dei vincoli | A |
| versione_corrente | Progressivo dell'ultima versione visualizzata | N |
| ultima_versione | Progressivo dell'ultima versione dei vincoli | N |
| data_progetto | Data della creazione del progetto | AAAA/MM/GG HH:MI |

### 2.5.2. Parametri

**Formato del file:** file delimitato da ";"

**Nome File:** savePar.csv

**Cartella:** Directory di lavoro

| Nome variabile | Descrizione | Tipo |
|---|---|---|
| minstrato | Numero minimo di unità per strato | N |
| maxiter | Numero massimo di iterazioni | N |
| maxiterChromy | Numero massimo di iterazioni (Chromy) | N |
| Epsilon | Epsilon Formato: 1e-11 | N |

## Bibliografia


BANKIER M.D. (1988), “Power Allocations: Determining Sample Size for Subnational Areas”, The American Statistician, Vol.42, pp.174-177.

BELLHOUSE D.R. (1984), “A Review of Optimal Designs in Survey Sampling“, *Canadian Journal of Statistics*, Vol.12, pp.53-65

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

CAUSEY B.D. (1983), “Computational Aspects of Optimal Allocation in Multivariate Stratified Sampling”, SIAM Journal of Scientific and Statistical Computing, Vol.4, pp. 322-329

CICCHITELLI G., HERZEL. A., MONTANARI G.E. (1992), “*Il Campionamento Statistico*”, Il Mulino.

COCHRAN W.G. (1977), “*Sampling Techniques”*, 3rd ed., Wiley, New York

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

DAYAL S. (1985), “Allocation of Sample Using Values of Auxiliary Characteristic”, *Journal of Statistical Planning and Inference*, Vol.11, pp.321-328.

DI GIUSEPPE R., GIAQUINTO P., PAGLIUCA D. - (2004), "MAUSS: un software generalizzato per risolvere il problema dell'allocazione campionaria nelle indagini Istat", Istat, Collana Contributi, n. 7/2004

FALORSI P.D., BALLIN M., SCEPI G., DE VITIIS C., (1998) “Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall’ISTAT”, *Statistica Applicata* Vol. 10, n.2

HIDIROGLOU M.A., LATOUCHE M., ARMSTRONG B., GOSSEN M. (1995), “Improving Survey Information Using Administrative Records: the Case of the the Canadian Employment Survey”, *Annual Research Conference,* Bureau of the Census, pp.171-197.

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

KOKAN A.R., KHAN S. (1967), “Optimum allocation in multivariate surveys: an analytical solution”, *Journal of the Royal Statistical Society B.,* No. 29, pp. 115-125.

KUHN H.W., TUCKER A.W. (1951), “Nonlinear Programming”, *Proceedings of II Berkley Symposium Mathematical Statistics and Probability*.

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

SIGMAN R.S., MONSOUR N.J. (1995), “Selecting Samples from List Frames of Businesses”, in Cox B.G., Binder D.A. Chinappa B.N., Christianson A., Colledge M.J., Kott P.S. (eds) *Business Survey Methods*, Willey, New York.

## Appendice: costruzione file input “strati” per MAUSS

In questa appendice mostriamo come sia possibile, utilizzando una funzione del package R “mauss” (utilizzato dal software presentato in questo manuale), costruire uno dei due input richiesti da MAUSS, quello relativo agli strati in cui è suddiviso il frame della popolazione di riferimento.

Per verificare la disponibilità del package “mauss”, in ambiente R occorre eseguire il comando:

```
> library(mauss)
```

In caso il package non sia stato installato, occorre prioritariamente procedere alla sua installazione, che peraltro è contestuale all’installazione del software MAUSS.

Per l’utilizzo della funzione `buildStrataDF`, che consente per l’appunto la costruzione del file “strata” di cui MAUSS necessita, si danno due possibilità:

1. il frame da cui verrà selezionato il campione oggetto di disegno contiene informazioni relative alle variabili target (le Y) dell’indagine (è questo il caso, ad esempio, di frame contenenti dati di censimento, oppure dati amministrativi);

2. il frame non contiene tali dati: sarà allora necessario calcolare, per ogni strato, delle stime relative a medie e scostamenti quadratici medi delle Y, ricorrendo a fonti diverse (ad esempio, una precedente ripetizione dell’indagine, oppure indagini diverse).

Nel seguito, esaminiamo entrambe le possibilità.

### *1. Disponibilità di informazioni dal frame di selezione*

In ambiente R deve essere definito un dataframe nominato “`frame`”, contenente, per ogni unità presente, le seguenti variabili:

1. un identificatore univoco dell’unità (nessuna restrizione sul nome, può essere "cod");
2. (opzionale) identificativo dello strato a cui appartiene l'unità;
3. (opzionale) i valori di m variabili ausiliarie (da X1 a Xm);
4. i valori delle p variabili target (Y1 a Yp);
5. il valore del dominio di interesse per il quale si vuole produrre stime (denominato "`domainvalue`").

Per esempio:

```
> frame <- read.delim(“frame.txt”)
> head(frame)
   cod domainvalue     strato X1 X2 X3        Y1        Y2
1  100           4 4so1b4sau1  2  4  1 3283.2128 1167.9092
2  200           4 4so1a6sau1  1  6  1 1997.4587  614.9569
3  300           4 4so1a6sau1  1  6  1  569.9164 1498.6392
4  400           4 4so1a8sau1  1  8  1 1786.8751 1051.1127
5  900           4 4so1a5sau1  1  5  1  910.3036  808.0705
6 1200           4 4so1b1sau2  2  1  2 3273.3433  969.6291
```

Se tali informazioni sono disponibili, è a questo punto immediato utilizzare la funzione `buildStrataDF` in questo modo:

```
> buildStrataDF(frame)
```

La funzione prende come argomento unico il nome del frame, e scrive nella directory di lavoro (con nome "strata.txt") il dataframe contenente informazioni sugli strati, così strutturato:

```
> head(strata)
  strato   N        M1        M2        S1        S2 cost cens DOM1 X1 X2 X3
1  1*1*1 156  623.4663  843.2696 469.92162 355.71351    1    0    1  1  1  1
2  1*1*2  68 1062.4884  867.4100 504.12793 366.40575    1    0    1  1  1  2
3  1*1*3  17  937.9182  905.4114 505.92665 327.92656    1    0    1  1  1  3
4  1*1*4  20 1377.0881  787.4087 359.69583 394.92049    1    0    2  1  1  4
5  1*1*5   3 1614.3787  660.2262  20.33451 250.12945    1    0    2  1  1  5
6  1*1*7   2 1809.0502 1324.6433 185.48919  86.84577    1    0    2  1  1  7
```

***2. Disponibilità di informazioni da fonti diverse dal frame (altre indagini)***

Al contrario, se non ci sono informazioni nel frame riguardanti le variabili target, è necessario costruire il dataframe "`strata`" a partire da altre fonti, per esempio una ripetizione precedente della stessa indagine, o da altre indagini. In questo caso, supponendo che le informazioni a disposizione siano contenuti in un file, abbiamo bisogno di leggere i dati eseguendo:

```
> samp <- read.delim("samplePrev.txt")
```

Oltre ai vincoli di denominazione introdotta in precedenza, questa funzione richiede che una variabile denominata "`weight`" sia presente nel dataframe "`samp`". E' a questo punto possibile eseguire la funzione nello stesso modo già visto in precedenza:

```
> buildStrataDF(samp)
```

Il risultato è praticamente identico a quello del caso precedente: la funzione scrive nella directory di lavoro il dataframe "`strata`", denominato "strata.txt". L'unica differenza è che le medie e gli scostamenti quadratici medi delle Y sono a questo punto delle stime campionarie, la cui affidabilità deve essere valutata attentamente considerandone la relativa varianza di campionamento

***Si noti che in tutti i casi, per ogni variabile target Y, media e deviazione standard sono calcolati escludendo i valori mancanti (NA)***