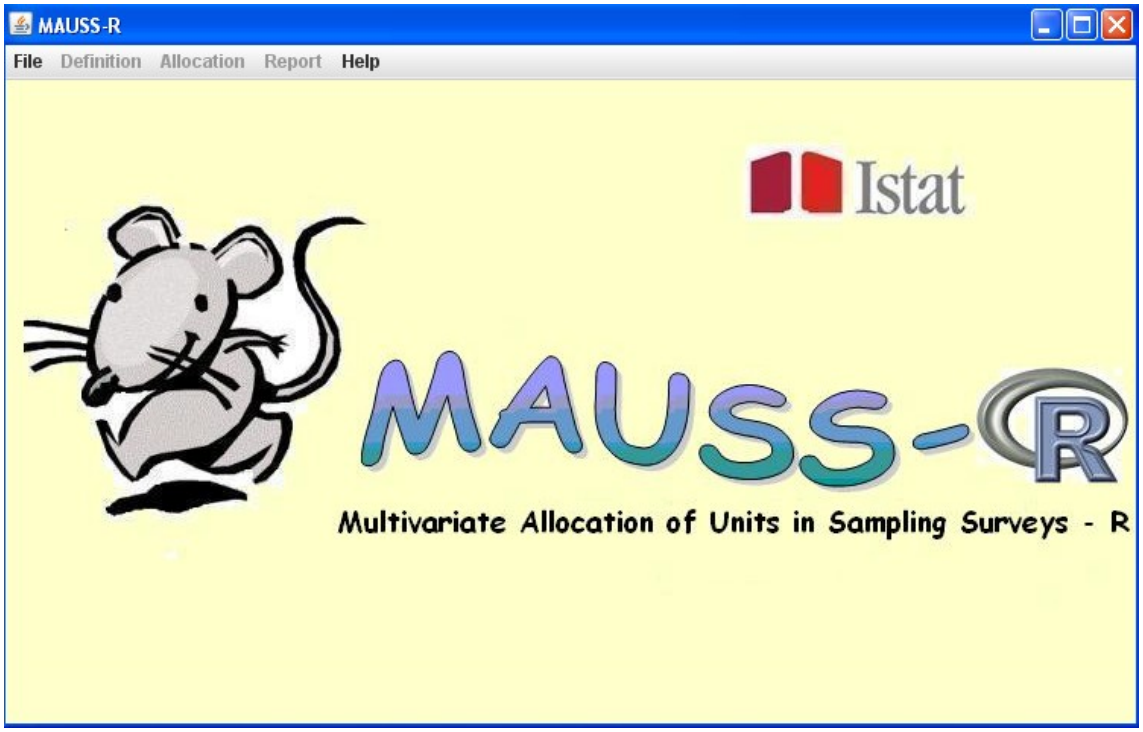
MAUSS-R

Multivariate Allocation of Units in Sampling Surveys



User and methodological manual

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

Introduction

Mauss is a tool for defining the sampling design for sample surveys on finite populations. It guarantees optimality criteria, flexibility and easy management for those who have the responsibility to design and conduct such surveys.

It enables the user, once defined the objectives and the operational constraints of the survey, to choose the best sampling design between those obtained by adopting different definitions of the key features of the survey, such as the type of stratification, the desired accuracy of the estimates, the sample size, the type of domains of study, the variables of interest.

The use of this software also ensures transparency, standardization and accuracy of the methods used.

The current version of Mauss is an evolution of previous applications, developed in SAS. The design and development of these first versions is due to methodologists and IT developers (including among the first Marco Ballin, Claudia De Vitiis, Piero Demetrio Falorsi, Germana Scepi; in the second, together with Marco Ballin and Piero Demetrio Falorsi, also Daniela Pagliuca, Paolo Floris and Roberto Di Giuseppe).

The decision to migrate the SAS version to R was taken as part of a strategy that tends to reduce the dependence on proprietary software and to ensure full portability of the tools developed by ISTAT. Moreover, new functions have been added, together with a more advanced interface. The development of the version described in this manual is due to Teresa Buglielli (Java interface for project management and execution modules), Daniela Pagliuca (implementation of the methodology in R) and Giulio Barcaroli (Chromy algorithm in Fortran).

1. The methodology implemented in MAUSS

1.1. Definition of the methodological problem of allocation

1.1.1. The planning of the survey sampling design

In designing a sample survey, the phase of studying the sampling design and defining the sample size and its allocation among strata requires the specification of a set of parameters and information, from which the construction of the input for the allocation procedure follows. It is necessary to determine:

- the population of interest
- the sampling unit
- the selection frame containing the unit of the population
- the variables of interest
- the parameters which are to be estimated,
- the level at which the estimates have to be produced, i.e. the domains of estimate
- the accuracy to be guaranteed for the estimates at the level of the different domains
- the auxiliary information useful for the design.

The *population of interest* must be defined on the basis of criteria that identify precisely the unit of analysis to be surveyed. Examples of populations are: the set of active enterprises in Italy with reference to a certain period of time, the population of households living in Italy in a fixed point in time, the babies born in Italy in a given calendar year.

The *selection frame* is the list of the units belonging to the population, containing at least the information required to identify and contact them. It may also contain auxiliary information useful for the design phase. In some cases, the frame identifies groups of units, or clusters, such as a list of families where the family is a cluster of individuals, or the register of Italian municipalities in which the town is a cluster of households.

The *variables* to be collected may be qualitative (qualitative answers to questions such as employment status or perception about a certain phenomenon) or quantitative (such as income, production or sales). Therefore the parameters to be estimated may be, as in the first case, the absolute or relative frequencies of response items, or, as in the second case, averages or totals. Anyway, the software considers as parameters to be estimated the totals of these variables, corresponding, therefore, to the absolute frequencies for qualitative items and to the totals for quantitative variables.

The *domains of estimates* are the sub-populations at the level of which the estimates of the parameters of interest have to be obtain. These domains must be defined on the basis of variables available in the frame for each the unit of the population. Examples of domains are: the region, the province, the region cross-classified with the economic activity (for enterprises), the age groups. In addition, it is often necessary that the estimates are produced for more than one type of domain, or to alternative partitions of the same population.

The *precision* required for the estimates of interest represents the degree of reliability that the estimates have to guarantee. It is expressed in terms of the coefficient of

variation (ratio between the standard error of the estimate and the estimate itself), to be specified for each parameter and each type of domain. For example, it is possible to require that the estimate of the total turnover of the enterprises at level of region presents a coefficient of variation not exceeding 10%. It is important to note that for a certain variable, the coefficient of variation is the same when considering the estimation of the average and of the total; for qualitative variables it is the same for the estimation of a relative frequency and the correspondent absolute frequency.

The *auxiliary information* useful for the planning of the design is generally contained in the frame or can be obtained from previous similar surveys or from a census. The auxiliary variables necessary for the allocation are: stratification variables, which are essential for defining strata and domains of estimate, variables correlated with the ones of interest, useful for the study of the variability of the variables of interest.

1.1.2. One-stage stratified sample design

MAUSS allows to calculate the sample size and its allocation in the strata for a one-stage stratified sample design. To accomplish this sampling scheme, the population should be divided into strata, accordingly to one or more classification variables known a priori for all units in the frame.

In a standard stratification, strata may be regarded as the minimum partition of the population that allows to obtain the domains of estimate as a union of strata (planned domains). In general, finer strata produce an increase of sample size, given the expected error; this is due to the necessity to ensure at least one or two sample units per stratum.

In order to illustrate a standard procedure for the construction of the strata, let's consider, for example, the case of a business survey aiming at producing estimates separately for classes of economic activity (as identified by the first four digits of the classification of economic activities, Nace) and size classes of employees. In this situation, the strata are defined by the cross-classification of economic activity and size class of employees.

The allocation of the sample size among strata is achieved following an approach which is a generalization of the method of Neyman (known as a method of univariate optimal allocation) and allows to minimize the sample size having established constraints on maximum expected sampling errors of target estimates, for each type of domain: we can define this approach as a multivariate and multi-domain allocation. The methodological aspects are described in detail in Section 5.

It is important to add that some strata can be defined as *take-all* strata on the basis of a decision of the responsible of the survey (for example, you may decide a priori to include in the sample all firms with more than 20 employees).

1.2. Preparation of input files

MAUSS requires that the user provides input data related to the characteristics of the population under investigation, to the variables of interest for the estimates, together with the constraints on the expected sampling error of the estimates.

As output, the system produces the sample size per stratum, the expected sampling errors of all estimates of interest and useful information to evaluate the solution found.

The input information must be provided to the software in two separate data files:

1. the first one contains the stratification of the population, with the number of units within each stratum, the indication of the domains of estimate and some estimates of the intensity and variability of the phenomena of interest;
2. the second one contains the constraints on sampling errors, specified for each variable of interest and each type of domain.

1.2.1. Strata file

The first file have to contain one record for each stratum with the following variables (rules on names and formats are given in chapter 2 of this manual):

- stratum identifier, *h (h = 1, .., H)*;
- number of units of the population belonging to stratum *h*, N_h;
- domain code of type 1, type 2, ..., type D to which the stratum *h* belongs;
- population means, calculated for each stratum and for each one of the P target variables that will be used to allocate the sample:

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

where Y_{pj} is the value of the variable y_p (p = 1 ,..., P) in the j-th unit of the population; for qualitative variables, you have to define a dichotomous variable for each response item, and the mean of the variable corresponds to the relative frequency $f_{p,h}$ of the value 1 of the dichotomous variable y_p :

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

where $F_{p,h}$ is the absolute frequency of the item;

- standard deviations of the P target variables in the population, calculated for each stratum:

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

for categorical variables, the standard deviation will be calculated as

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

- indication on stratum to be sampled or taken-all (0 to be sampled, 1 otherwise);
- fieldwork costs in the stratum (cost per each interview).

For the construction of the first data set, the main difficulty may arise from the obtaining of auxiliary information on the variables of interest.

There may be given different possible situations:

3. means and standard deviations can be inferred from the sampling frame, but are referred to a previous time reference and / or to proxy variables of the variables investigated;

4. means and standard deviations are obtained as estimates from a previous occasion of the same sample survey;
5. means and variances are unknown.

In the first case, which occurs for example in the situation of a business survey, when the turnover or the number of employees is available from the frame (a business register) for each enterprise referred to a previous year, it is immediate to calculate for each stratum the required quantities, according to expressions (1a) (1b), (2a) and (2b). Often, the variables available on the frame are "proxy" of the variables under investigation and if the correlation between the auxiliary variables and the variables of interest is high enough, it is possible to ensure a good level of precision on the estimates of the variables of interest (Cicchitelli *et al*., 1992).

In the second case, it is possible to obtain the estimates of means and standard deviations in the population, from sample data of a previous occasion of the same survey. In this case it is necessary, however, to evaluate the reliability of these estimates and to use them at a higher level of aggregation than the stratum if they do not exhibit an acceptable accuracy. In the following, we will show an example dealing with this particular situation.

The third situation happens when the user does not have any information at all on the variability of the phenomena of interest because the survey is planned for the first time. In these cases, it is possible to set the allocation procedure by establishing, for each domain of estimation, a set of "typical" frequency estimates, in order to cover the range of variation for all estimates that the survey aims to produce. For instance, if the survey is used to produce estimates at national, macro-regional and regional levels, you might desire that the sample is such as to guarantee a sufficient reliability for estimates at least of 1% at national level, 3% at macro-regional level, and 5% at the regional level. In this case, the strata will coincide with the most disaggregated domain, namely the region, using three variables whose means will be constant for all strata:

$P_{1,h} = 0.01,\ P_{2,h} = 0.03,\ P_{3,h} = 0.05$ for each stratum h,

while the standard deviations can be obtained using the (2b).

The software will provide the overall sample size and its allocation among the strata in such a way that the constraints are respected with regards to sampling error of the estimates of "typical" frequencies at the level of the different specified domains.

1.2.2. File of constraints on maximum expected sampling errors

The second file should contain one record for each type of domain with the following variables:

1. code of domain type, *d (d = 1, .., D)*;
2. maximum allowable values of the expected coefficients of variation for each one of the total estimates of K variables of interest, CV1,... ,CVK.

The preparation of the second file requires the user to specify for each of the estimates of interest the maximum value of the coefficient of variation allowed for each type of domains.

It is worth noting that if for a certain estimate it is not needed to guarantee a limit for the sampling error for a certain type of domain, it is possible to indicate a very high value of the coefficient of variation for that type of domain, such as, for example, cv = 1.

Regarding the criteria used to set the level of error in the domain, it is common practice to allocate the sample so that the level is approximately equal for all domains (Sigman and Monsour, 1995). .

1.2.3. Example of construction of input data sets

The following example is based on the stratification adopted in ISTAT for the survey on births. The target population consists of mothers of babies born in a given year, stratified by age groups (5) and regions (21); the interview is conducted two years after the birth.
The estimation domains are region, macro-region, age group and nation. We assume for simplicity that the estimates of interest are only two: the relative frequency of women who were employed before the birth but no longer employed at the time of the interview, and the relative frequency of women whose children attend the nursery. Information on the variables of interest in this case is drawn from the data of an previous survey (case 2 in paragraph 1.2.1). This information was not reliable enough to be used at stratum level, but only by considering domains defined as cross-classification of macro-regions and age groups. The variable *cost* of each stratum is set equal to one because there is no difference in cost between the different strata. The same for the variable that indicates the presence of strata to be taken-all: in this survey this indicator has always been set to zero.

The resulting file has the following structure.

STRATUM	**Domain 1 = Region**	**Domain 2 = Age group**	**Domain 3 = Macro-region**	**Dom 4 = Nation**	**Pop**	**Mean 1**	**Std dev 1**	**Mean 2**	**Std dev 2**	**Cost**	**Cens**
15-24 Piemonte	Piemonte	15-24	North West	1	1000	0.20	0.4	0.45	0.497	1	0
25-29 Piemonte	Piemonte	25-29	North West	1	1600	0.18	0.348	0.5	0.5	1	0
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
35-39 Sardegna	Sardegna	35-39	Islands	1	700	0.30	0,458	0.2	0.4	1	0
40 → + Sardegna	Sardegna	40 e oltre	Islands	1	300	0.30	0,458	0.25	0.433	1	0

The second file, containing the constraints on sampling errors, has the following structure.

Domain type	CV1	CV2
Dom1 = Region	0.10	0.14
Dom2 = Age group	0.08	1
Dom3 = Macro-region	0.05	0.08
Dom4 = Italy	0.02	0.03

The values assigned to the coefficients of variation for the two estimates at level of the four types of domains are only examples, but show how, in general, to the types of domains with a larger number of values, is given a higher value of the coefficient of variation.

It can be noted that, as for the second variable is not required to estimate the level of age groups (DOM2), the bound was set equal to 1.

1.3. How to use the output of the system

The system produces as output: (a) the sample size per stratum, (b) the expected sampling error for each target estimate in each domain of interest, (c) some useful statistics for the improvement of the sampling plan.

The sample sizes for each stratum are added to the input dataset of strata, while the expected sampling errors are reported both in the output dataset and in output tables 7 and 8. The statistics useful for adjusting the allocation solution are shown in Table 5. The system allows the user to choose the final solution by comparing the results of several tests, obtained by defining the precision constraints in different ways.

Table 5 is the instrument at the user's disposal to evaluate how to modify input data, particularly data in the second input file. This table contains the information useful for the sensitivity analysis: for each estimate and each type of domain is given the value of the additional sample size needed to achieve a decrease of 10% of the coefficient of variation of the corresponding estimate. This number can also be interpreted in the opposite direction, i.e. as it represents the decrease in sample size that would be achieved by increasing the error of the corresponding estimate of 10% at the level of that type of domain.

For example, continuing the example regarding the survey on births, suppose that the sensitivity of the estimate of the first variable in the first domain type (region) is equal to 567 units. Because the coefficient of variation of this estimate was set at 10% (CV1 = 0.10), this means that:

- to obtain a reduction in sample size of 567 units is necessary that the value of CV1 shifts from 0.10 to 0.11, which is equivalent to an increase of 10% of the expected sampling error;
- to obtain a reduction of 10% in the expected error, 567 units should be added to the sample.

Using this tool the user is able to make the necessary adjustments to achieve the desired sample size or, conversely, to achieve the desired expected precision on target estimates.

1.4. The multivariate and multi-domain allocation methodology

In general, the determination of the sample size of the different strata is functional to the minimization of the sample variability of the estimates. In the absence of specific information on the variability in the strata, the objective is achieved through the proportional allocation; conversely, if this information is available, it is possible to define more efficient allocations.

In the case of a single variable of interest, being available an estimate of the variability, one can refer to well-known results for the optimal allocation in the univariate case (Cochran, 1977); these results are used to determine the sample size with the aim to minimize variance estimation for a fixed value of the cost function or, conversely, to minimize costs, having previously established the level of accuracy of the estimates. The univariate solution is however not suitable for the design of most surveys, which are usually characterized by a plurality of target estimates. For these surveys, therefore, it is necessary to deal with the problem of optimal allocation under a multivariate approach. The following is taken from Falorsi *et al*. (1998).

1.4.1. Multivariate allocation

In a stratified sample with equal probabilities of selection of units and without replacement, the variance of the estimator of the total of a generic variable of interest, y_p (p = 1 ,..., P) can be expressed as:

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

where $V_p' = V_{p0} + V_p$ is the variance of variable p in stratum h and V_{0p} is the part of variance not influenced by allocation.

We also define the following cost function:

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

where C_0 is the fixed cost of interviewing that does not depends on the sample size nor on the allocation, C is the variable cost, and C_h (h = 1 ,..., H) the cost per sample unit in the stratum h.

It is possible to determine the number of units to be assigned to each stratum using two approaches (Sigman and Monsour, 1995). The first approach consists in minimizing the product $W * C$, where $W = \sum_{p=1}^{P} W_p V_p$ and W_p (p=1,...,P) are weights to be defined. The solution is found by setting the value of W or C. It is possible that this method does not work in concrete situations due to the difficulty of specifying non-arbitrary weights.

In the second approach an upper bound V_p^* is set for each V_p' and the cost function C is minimized under the constraints $V_p' \leq V_p^*$ (p=1,..,P).

MAUSS uses the latter approach, adopting a generalization of the solution proposed by Bethel (1989), that defines a constrained minimum problem with convex objective function and linear constraints. In particular, we reformulate the quantity C in (4) by defining:

$$x_h = \begin{cases} 1/n_h & \text{if } n_h \geq 1 \\ \infty, & \text{otherwise} \end{cases}$$

In this way the expression of the objective function to be minimized becomes:

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

where $\boldsymbol{x} = (x_1, \ldots, x_H)'$. The constraints $V_p' \leq V_p^*$ take the form:

$$\sum_{h=1}^{H} a_{p,h} x_h \leq 1 \quad , \; p=1,\ldots,P \tag{6}$$

being:

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

Since the minimization problem of (5) under constraints (6) satisfies the conditions of the theorem of Kokan and Khan (1967), then an optimal solution $\boldsymbol{x}^*$ exists. Using the theorem of Kuhn-Tucker (1951), Bethel demonstrates that there exist values $\lambda_p^* \geq 0$, so that the optimal solution takes the form:

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

where $\mu_p^* = \lambda_p^* \Big/ \sum_{p=1}^{P} \lambda_p^*$, therefore $\sum_{p=1}^{P} \mu_p^* = 1$ (9)

To determine simultaneously the optimal values x_h^* and μ_p^* it is necessary to resort to numerical algorithms, such as those proposed in the work of Bethel, which will be discussed in the following.

1.4.2. Multivariate allocation for multiple domains and multiple domain types

The solution described in the previous paragraph is related to the case when the estimates of the parameter of interest have to be provided for the total population. In

general, however, sample surveys are intended to provide estimates not only for the entire population, but also for subpopulations (domains of study) identified by a partition (or *domain type*) of the population under investigation. Furthermore, it is often necessary that the estimates are produced for more than one type of domain, which identify alternative partitions of the same population. In these cases the sample must be planned so as to ensure simultaneously the accuracy of the estimates at different required levels of detail, and this can be achieved by generalizing the solution previously described.

To illustrate the method of multivariate allocation in the case of multi-domain estimation, we denote by d (d = 1 ,..., D) the generic type of domain; k_d (k_d=1,...,K_d), the generic domain of type d; H_{k_d} the number of strata belonging to the domain k_d. The objective function (5) remains unchanged, while the system of constraints can be redefined as follows:

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

where V_{p,k_d}^* is the upper bound on the sampling variance of the estimate of the total of variable p for the domain k_d.

Similarly to what was done in previous paragraph, the (10) can be written as:

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

where

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

with $\delta_{k_d,h} = \begin{cases} 1 & \text{if } h \in k_d \\ 0 & \text{otherwise} \end{cases}$

By defining an index r whose values are in correspondence with the values found by lexicographically ordering the vector identified by three indices (d, k_d, p), the system of constraints becomes:

$$\sum_{h=1}^{H} a_{r,h} x_h \le 1 \qquad \text{for } r=1,...,R, \text{ where } R = P\sum_{d=1}^{D} K_d\ , \qquad (12)$$

i.e. a form totally equivalent to (6).

Returning to the (8), and being the conditions of the theorems of Kokan and Khan Kuhn-Tucker still satisfied, the optimal solution that minimizes (5) under constraints (12) is:

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

$$\text{where } \mu_r^* = \lambda_r^* \Big/ \sum_{r=1}^{R} \lambda_r^* \quad \text{with} \quad \sum_{r=1}^{R} \mu_r^* = 1. \tag{14}$$

1.4.3. Resolution algorithms

The algorithm proposed by Bethel for the calculation of the optimal multivariate allocation can be generalized to solve the same problem when there are multiple types of domains. This algorithm comes to the optimal solution iteratively, starting from an initial one ($\nu = 1$) which coincides with the optimal solution in the univariate case for the first variable on the first domain ($r = 1$). Typically with this solution the objective function assumes a very small value and the remaining constraints ($r = 2, .., R$) are not satisfied. In each of the following steps ($\nu = 2.3,\ldots, \nu$), the sample size is increased, increasing the objective function $f\left(x^{(\nu)}\right) \geq f\left(x^{(\nu-1)}\right)$ in order to satisfy all the constraints. Bethel shows that the algorithm converges and, therefore, μ^* e x^* can be identified simultaneously so that $0 \leq f\left(x^{(\nu)}\right) \leq f(x^*)$.

The computational complexity of this algorithm, especially in the case of multiple domains of study, led to the use of the algorithm proposed by Chromy (1987) which is of more immediate implementation and seems to converge towards the optimal solution more quickly.
To illustrate this algorithm, let $\boldsymbol{A}=\{a_{r,h}\}$ be the matrix of size R and H, whose elements are defined by (11) and $\boldsymbol{a}_r$ be the r-th row of A. The Chromy algorithm is an iterative algorithm, whose first step consists in computing the value of x according to (13), by setting each element of μ equal to $1/R$. If this solution satisfies all the constraints, the algorithm stops. Otherwise, the algorithm calculates $\boldsymbol{x}^{(\nu)}$ in correspondence of vector $\boldsymbol{\mu}^{(\nu)}$ whose generic element is provided by the following expression

$$\mu_r^{(\nu)} = \mu_r^{(\nu-1)} \left(a_r x\left(\mu^{(\nu-1)}\right) \right)^2 \Big/ \sum_{r=1}^{R} \mu_r^{(\nu-1)} \left(a_r x\left(\mu^{(\nu-1)}\right) \right)^2 \qquad 1 \leq r \leq R \tag{15}$$

where $x\left(\mu^{(\nu-1)}\right)$ denotes the value of x, obtained on the basis of (13) putting $\mu = \mu^{(\nu-1)}$.

Since these algorithms do not ensure that in the optimal solution satisfies $n_h \leq N_h$, MAUSS contains a procedure for the iterative reallocation that sets as take-all strata the strata in which $n_h > N_h$ and recalculate the sample size under the changed conditions.

2. MAUSS: user manual

2.1 Installation

Microsoft Windows.

The minimum hardware requirements for Mauss-R are:

- RAM: 512MB
- Disk Space: 5MB

Also need to be installed on your PC:

- Java 2 Runtime Environment version 6 or higher (http://java.sun.com/javase/downloads/index.jsp)
- R Environment version 7.0 or higher (http://cran.r-project.org/bin/windows/base/)

The environment variable PATH must point at the programs *java.exe* and *r.exe*.

To change the variable PATH:

Start → Settings → Control Panel → System → Advanced → Environment variables

Now select the PATH variable and click on the *Edit* button. Add here, at the beginning of the string, the path to the folder that contains the *java.exe* file and the folder that contains *r.exe* separated by ";".

For example:

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;

Installation

To install the software you need to download the file setup_MaussR.exe on your PC and run it.

2.2 Use of the software

2.2.1. Starting MAUSS

From Windows Menu:

Start ->Programmi->mauss->MaussR

From the desktop: double-click on the icon 

2.2.2. Main menu

The MAUSS-R menu contains the following functions (see fig. 1):

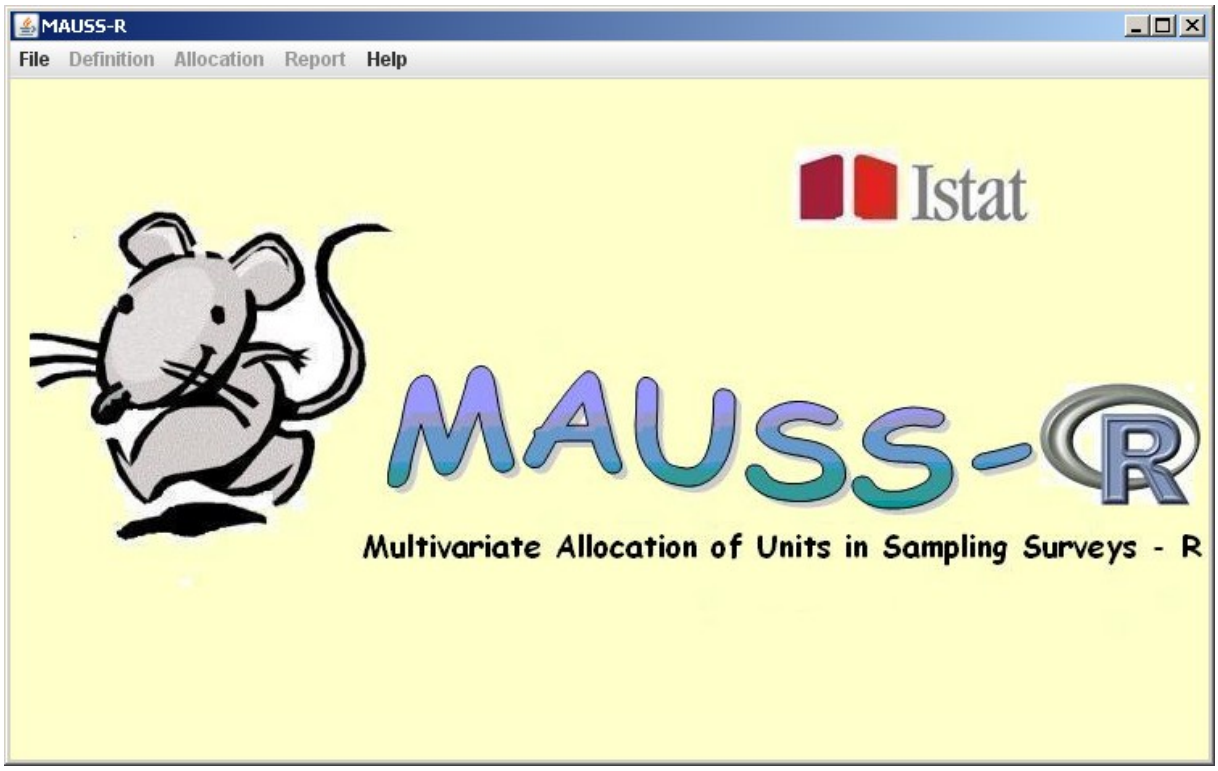


Figure 1 – Main Menu

- *File* – Definition of a project: creation of a new project, opening an existing project, closure of the project in progress and quit the application.
- *Definition* – Changing the parameters and the constraints used to compute the optimal allocation.
- *Allocation* – Running the optimal allocation by the method of Bethel for the current or for all versions of constraints file.
- *Report* - View the results and prints.
- *Help* - Display the online help.

2.2.3. Project definition

In MAUSS-R, a "project" is individuated by the name of the folder in which all data files generated by the application will be located. Other relevant information are the names of input files prepared by the user:

1. the first one gives the population size and mean and variance for each variable of interest for each stratum.
2. the second one includes, for each domain, the coefficients of variation for the estimates. For a description of the two files see below the section "Data file description".

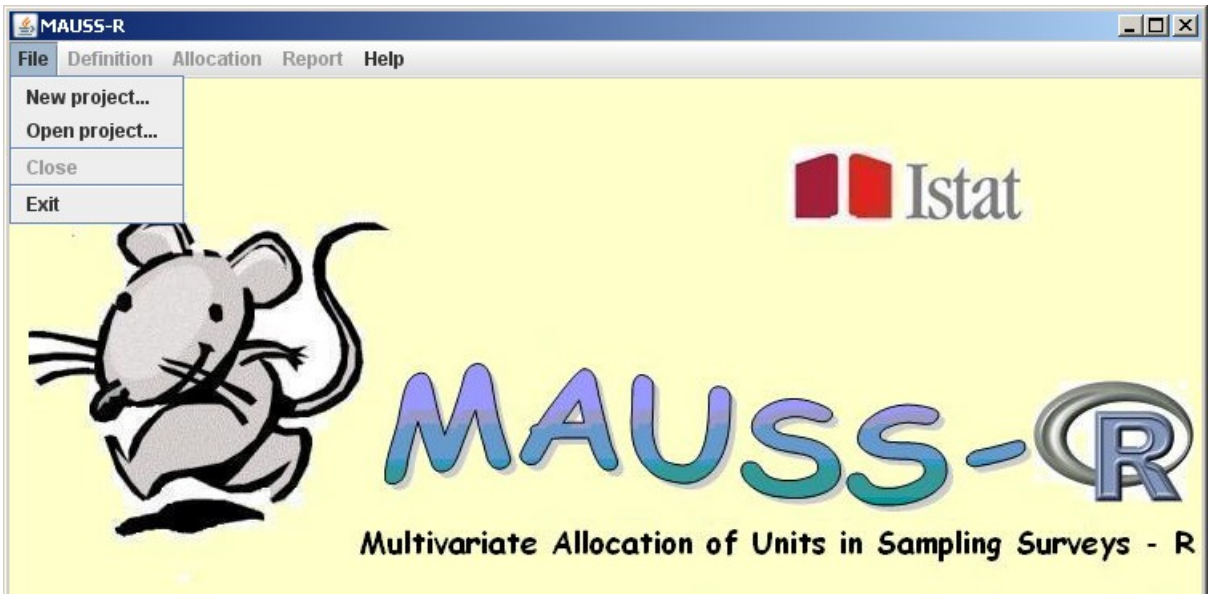


Figure 2 – File menu

Functions (see fig. 2):

- **New project**: : Inserting a new project.

Choosing the item New project, the window shown in Figure 3 will be open. This window allows to choose the folder in which the result files will be written and the two input files.

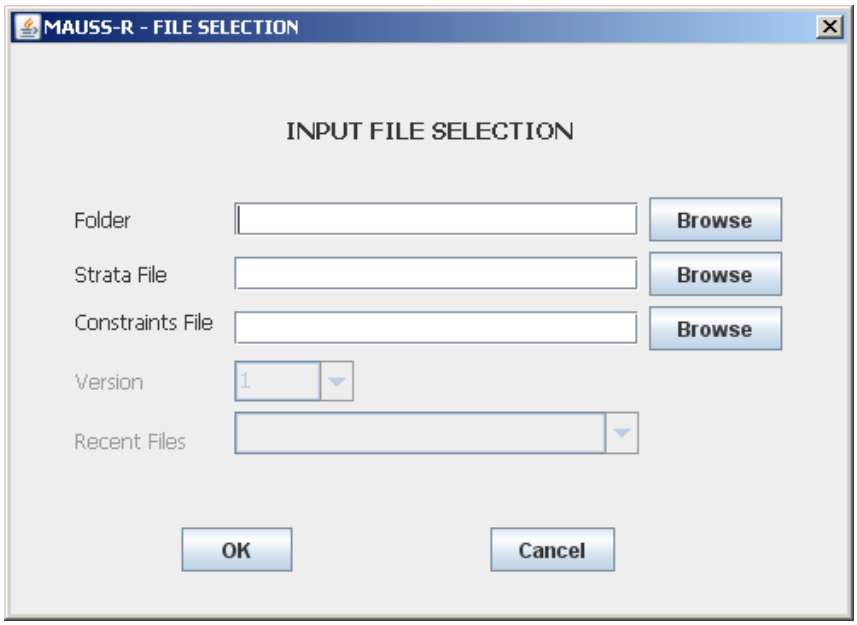


Figure 3 - New project

File names may be entered directly into the text box or can be selected using the File Manager clicking on the Browse button.

After giving the confirmation (OK), the procedure checks the data entered and, if everything is right, prepares the environment: sets the version number of the constraints to 1 and creates the BethV1 subdirectory of the work folder where copies the constraints file and where will write the results of the optimal allocation for the first version of constraints.

If an old project was defined in the chosen folder, the system asks if you want create a new project. If so, it cleans the folder by moving all the results of prior process in a sub-folder named backupNNNNNN where NNNNNN is a number that represents the system time in milliseconds. Otherwise it closes the window without defining the project that can be opened using the Open Project function.

- **Open project**: Opening an existing project

In this case, the user can choose the version of the file of the constraints and the project, from a list of already defined projects. The fields for the choice of the folder and of the two input files will automatically written and can not be changed.

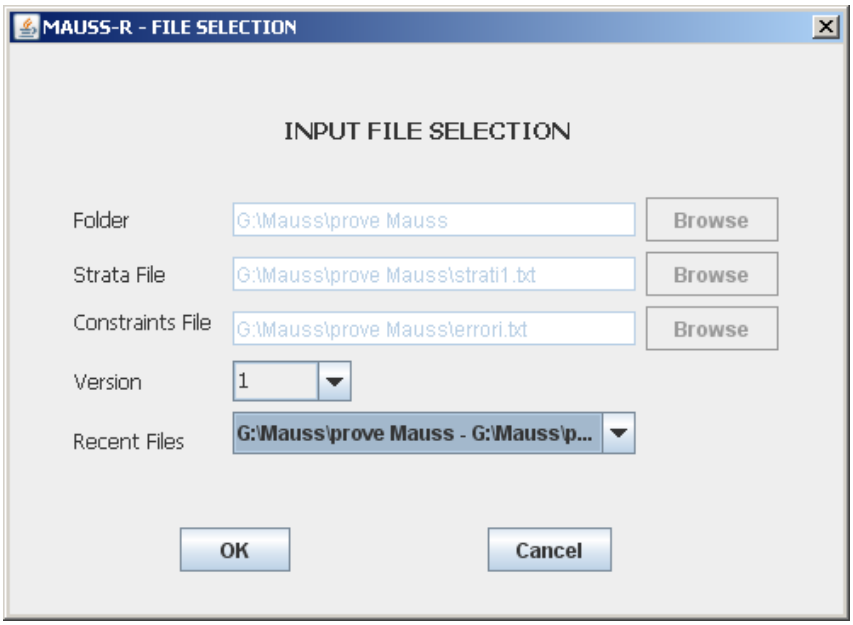


Figure 4 - Open Project

- **Close**: Closing the current project.
- **Exit**: Quit the application.

2.2.4. Parameters and constraints definition

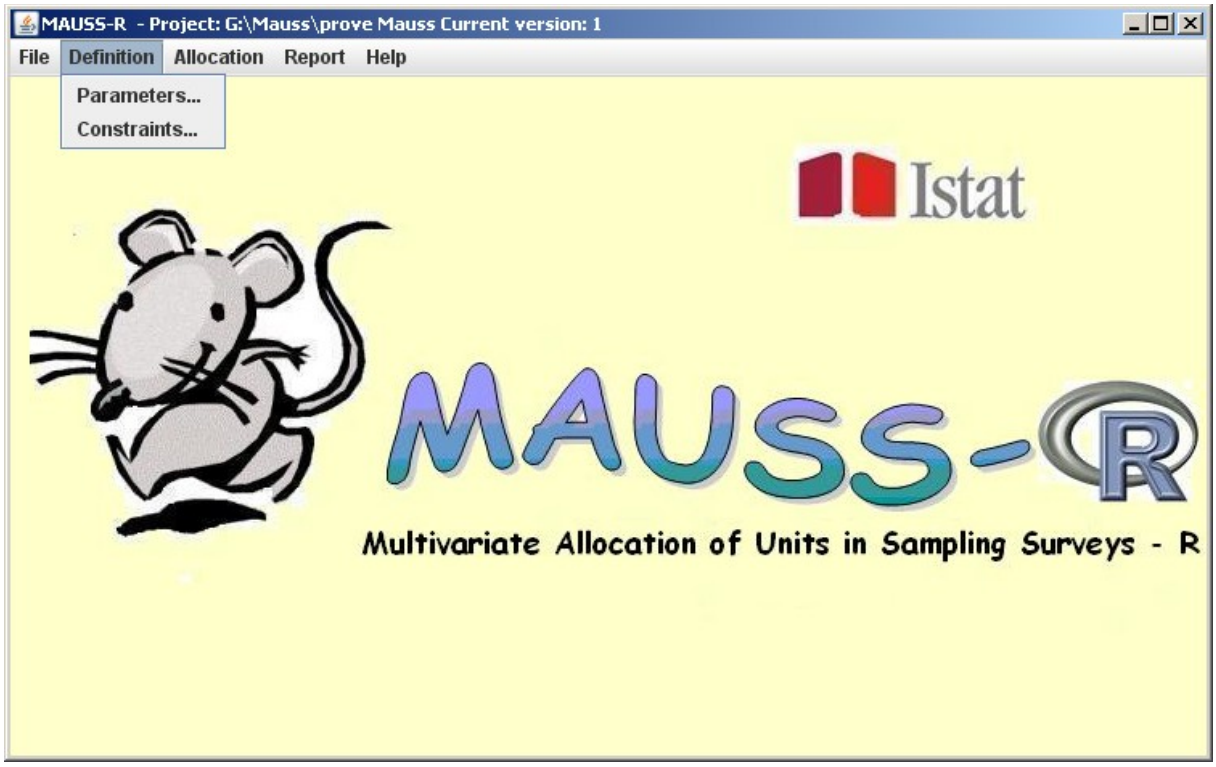


Figure 5 Definition Menu

Functions (see fig. 5):

- **Parameters:** Definition of parameters.

It is possible to modify the following parameters (see fig. 6):

- Minimum number of units per strata (default = 2).
- Maximum number of iterations (default = 25) of the general procedure. This kind of iteration may be required by the fact that when in a stratum the number of allocated units is greater or equal to its population, that stratum is set as "census stratum", and the whole procedure is re-initialised.
- Maximum number of iteration in the algorithm of Chromy (default = 200);
- Epsilon (default = 1e-11): this value is used to compare the difference in results from one iteration to the other; if it it is lower than "epsilon", then the procedure stops.

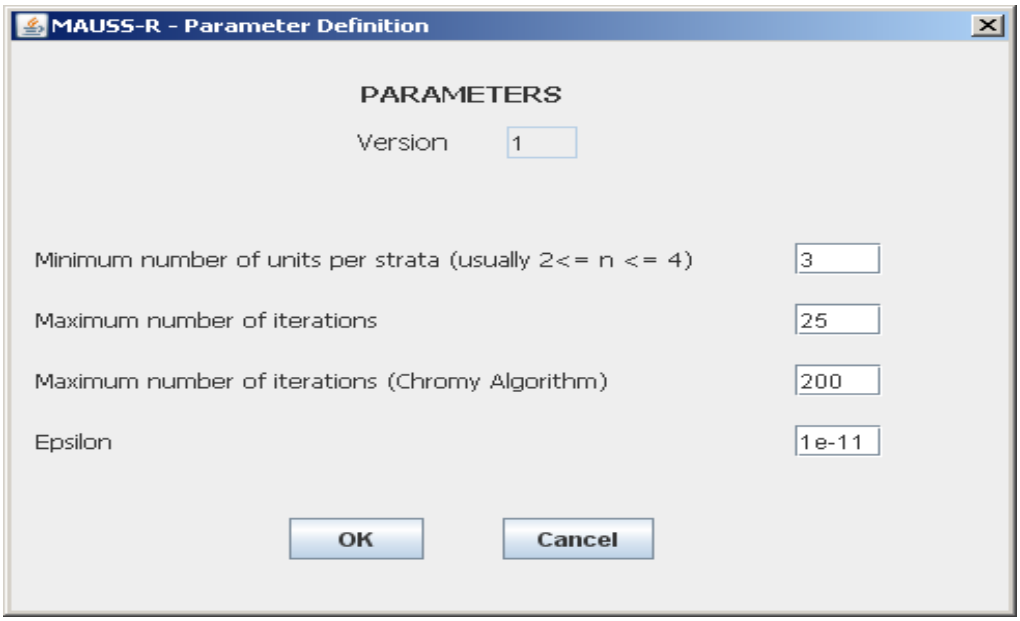


Figure 6 – Parameter Definition

- **Constraints:** Definition of constraints.

This function allows (see fig. 7):

- Choose the version of constraints.
- Modify the values of constraints in the table.
- Insert a new version of the file.

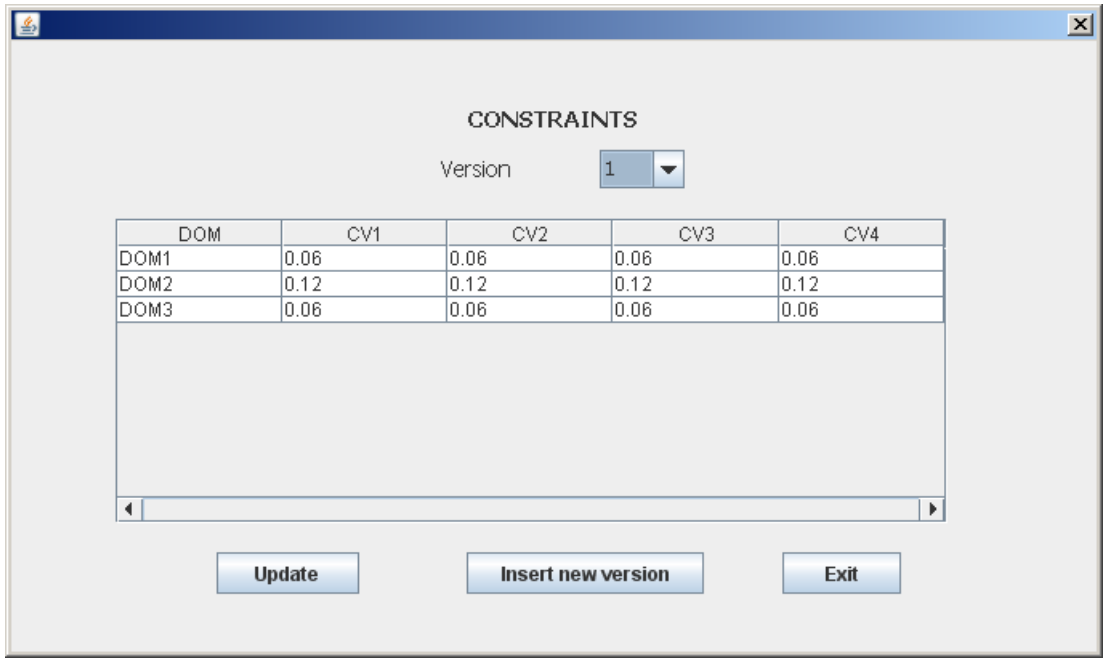


Figure 7 - Definition of constraints

You can change a coefficient of variation by double-clicking the cell, writing the new value and moving to the next box with the TAB key or the mouse.

WARNING! The change is recorded only if the cursor is positioned in a cell different from the cell that contains the changed value.

The *Update* button writes the constraint's table on the current version of the file.

The button *Insert* changes the version number of the constraints file, creates a new sub-folder of the working directory with the name BethVn, where n is the new version number, and inserts the data displayed in the table to a new file of constraints.

It is possible change the current version of the current constraints file using the list-box *Version.*

2.2.5. Allocation

This menu runs the function that computes the multivariate optimal allocation for different domains of interest in a stratified sample design. This function is an extension of Bethel methodology with Chromy Algorithm.

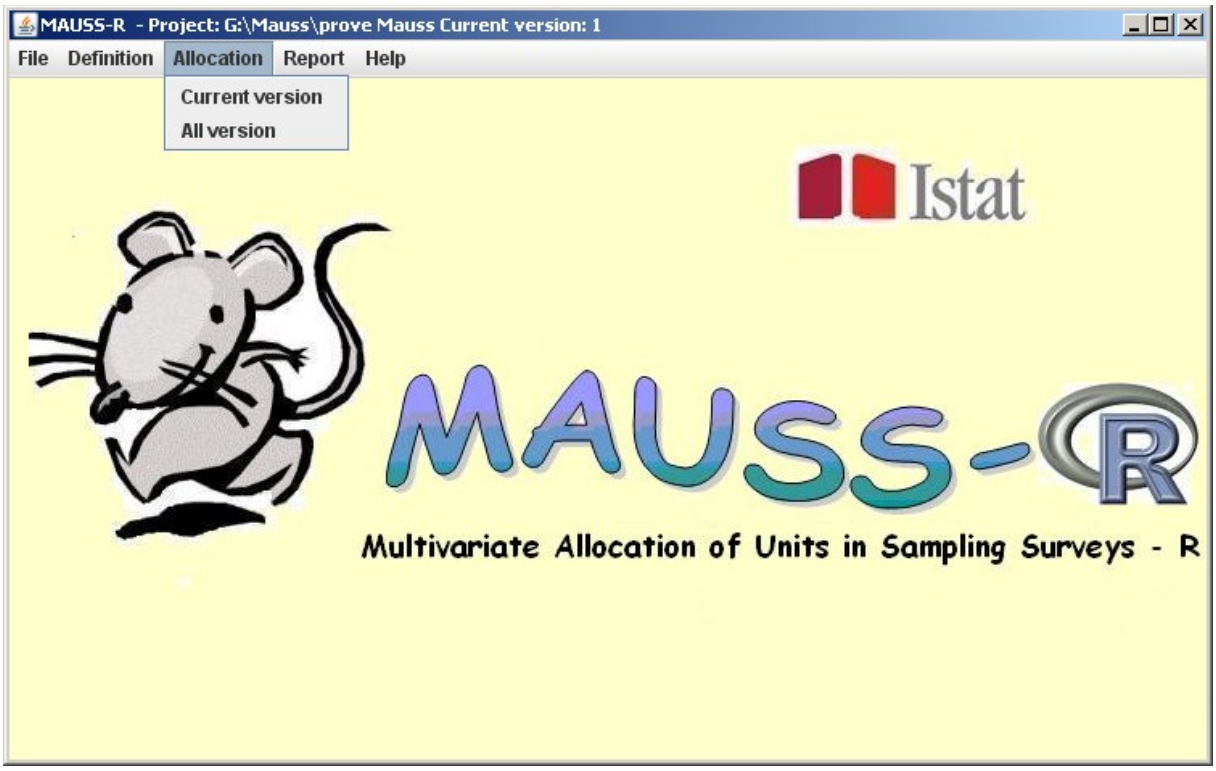


Figure 8 - Allocation Menu

Functions (see fig. 8)

- **Current version**: Calculating optimal allocation for the current version.
- **All versions**: Calculating optimal allocation for all versions of the file of constraints.

2.2.6. Reports

Reports concerning general information about population and results of Bethel method are available.

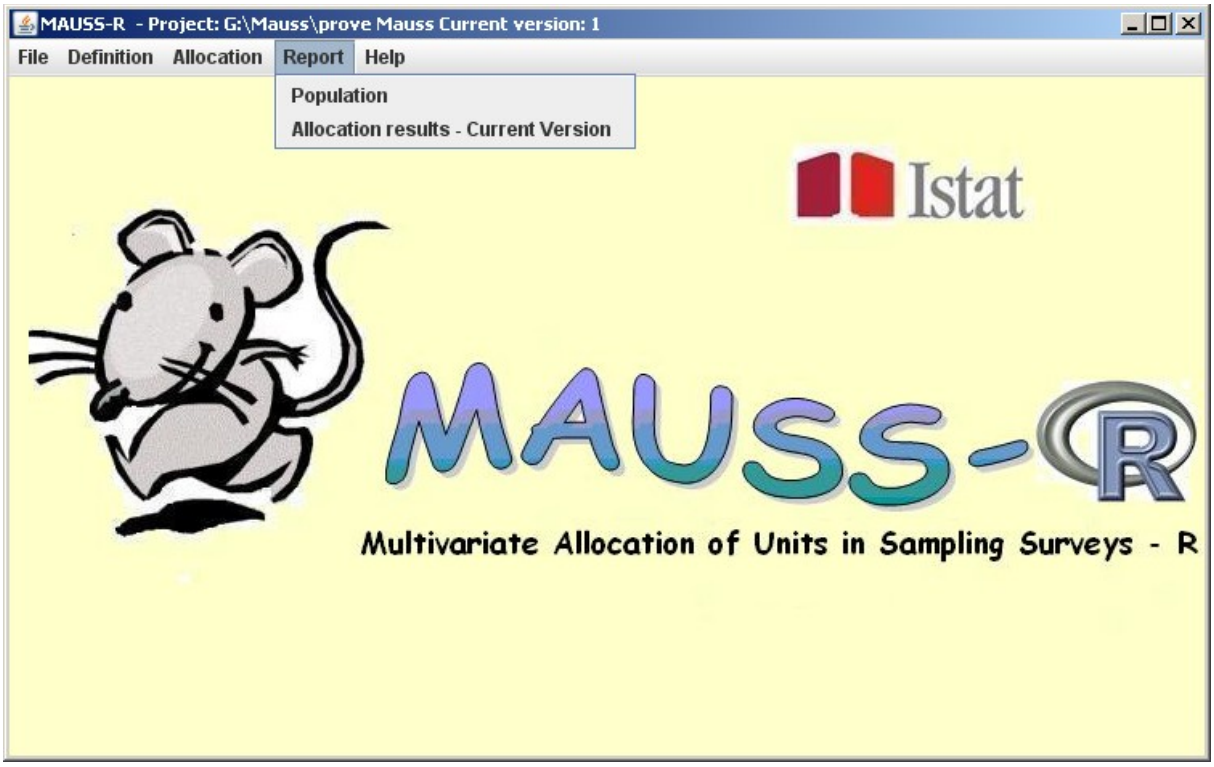


Figure 9 – Report Menu

Functions:

- **Population**: Analysis of the population.

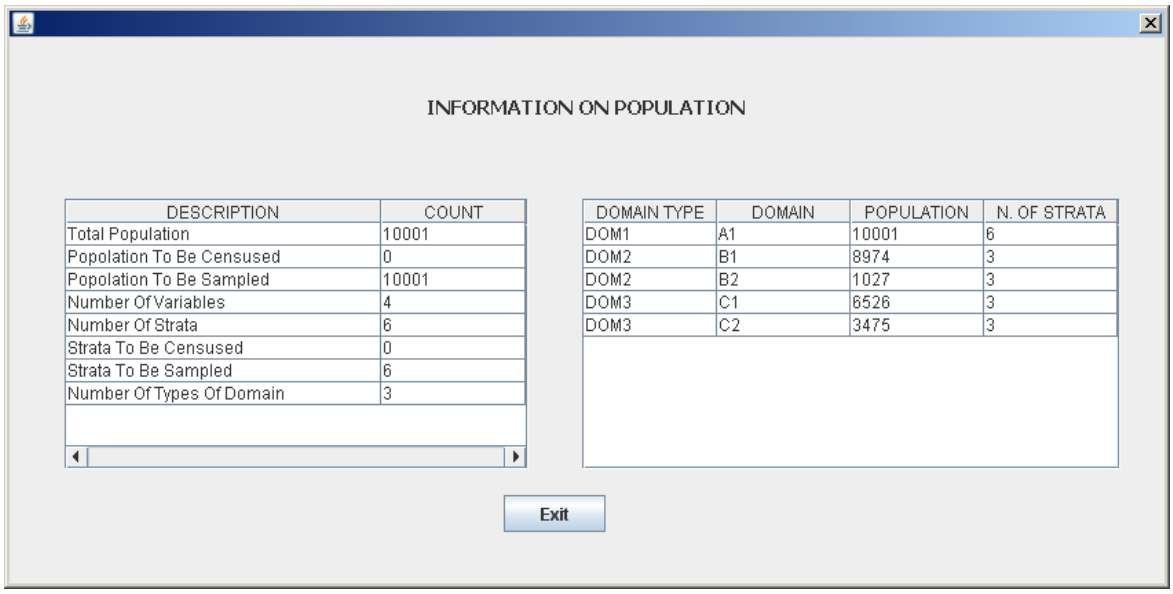


Figure 10 – Information on population

In this window (see fig. 10) there are two tables containing information about population. In the first there are general information as the number of population units, the number of strata and of different domains. The second is a table of population and number of strata by domain for every kind of domain.

These reports are written in *Bethel_Report_Pop1.xls* file.

- **Allocation Results**

In this window (see fig. 11) there are three tables containing information about optimal allocation.

Comparison between allocation results table: for each stratum, the sample size obtained by the optimal allocation of Bethel for the different versions of the file of constraints. This report is written in *BethelResults.xls* file.

Allocation results table: for each, stratum, the Bethel sample size, computed with the current version of coefficient of variation, is compared with the dimension of population and the values obtained with proportional and equal allocation.

This report is recorded in the *Bethel_Report1.xls* file.

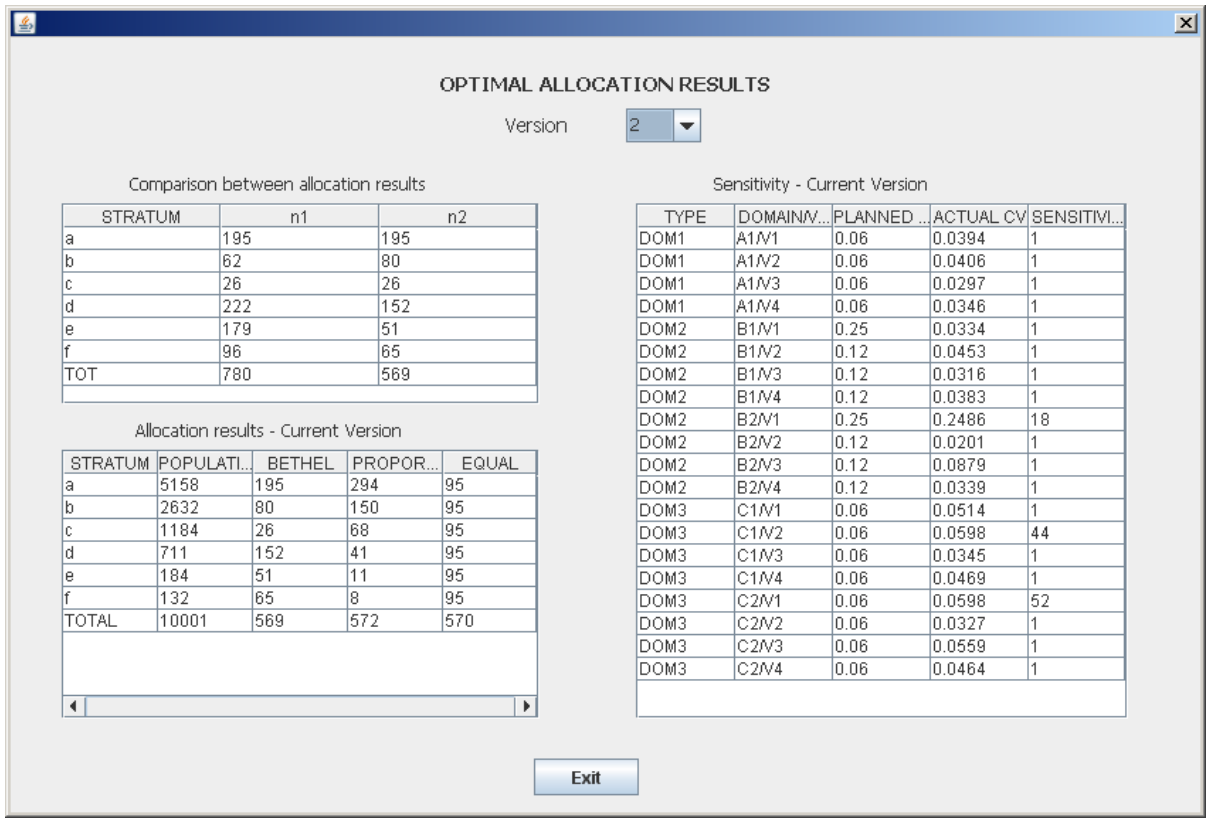


Figure 11 – Allocation results

Sensitivity: Coefficient of variation and sensitivity

In this table, expected and actual coefficients of variations and sensitivity to a variation of 10% of the desired precision are reported.

This report is recorded in the *Bethel_Report2.xls* file.

2.3. Input data

2.3.1. Strata file

File Format: tab-delimited (.txt).

Header: The first line of the file must contain the names of the variables specified in the table in any order. The file can also contain other variables.

Data: A record for each stratum with, at least, the information listed in the table (variables COST and CENS can be omitted). Data may also be related to other variables not involved in calculating optimal allocation.

Variable name	Description	Format
STRATO	Stratum code.	A
N	Number of population units in the stratum.	N
DOM1, DOM2, …,DOMp	Domain codes (1…p)	A
M1, M2, …, Mn	Means of *n* variables in the population	N
S1, S2, …, Sn	Standard deviation of *n* variables in the population	N
COST	Survey unitary cost for stratum. Default=1	N
CENS	Indicator of stratum coverage: 1 = stratum to be censused. 0 = stratum to be sampled. Default = 0.	N

2.3.2. Constraints file

File Format: tab-delimited (.txt).

Header: The first line of the file must contain the names of the variables specified in the table in any order. The file can also contain other variables.

Data: Coefficients of variation for all domains. A record for each type of domain that contains the information about the variables listed in the table. Other not relevant variables can be present.

Variable name	Description	Format
DOM	Type of domain code.	A
CV1, CV2, …, CVn	Planned coefficient of variation for n variables.	N

2.4. Produced output

OUTPUT FILE

File Format: tab-delimited (.txt).
Filename: *Bethel_campio.txt*
Folder: Sub-folder in the working directory for the version.

It is a copy of the strata input file to which is appended the variable CAMP containing the result of Bethel's optimal allocation.

2.5. Work datasets

2.5.1. List of projects

File Format: tab-delimited (.txt).
Filename: *progetti.txt*
Folder: Sub-folder $HOME/.Mauss2 in the working directory for the version.

Variable name	Description	Type
folder	Working directory	A
strati	Name of strata file	A
vincoli	Name of constraints file	A
versione_corrente	Progressive number of the last used version of the project	N
ultima_versione	Progressive number of the last used version of the constraints	N
data_progetto	Date of creation of the project	YYYY/MM/DD HH:MI

2.5.2. Parameters

File format: delimited file (with ";" delimiter character)

Filename: savePar.csv

Folder: Working **d**irectory

Variable name	Description	Type
minstrato	Minimum number of units per stratum	N
maxiter	Maximum number of iterations	N
maxiterChromy	Maximum number of iterations (Chromy)	N
Epsilon	Format: 1e-11	N

References

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

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.

Appendix: building input file “strata” for MAUSS

In this appendix we show how it is possible, using a function of the R package "Mauss" (used by the software presented in this manual), to build one of the inputs required by Mauss, the one relating to the strata that characterise the frame of reference population . To check the availability of the package "Mauss", in R environment you must run the command:

```
> library(mauss)
```

If the package has not been installed (but it should, as it is contextual to the MAUSS software installation), you must install it as a priority.

To use the function `buildStrataDF`, which allows for the construction of the input file "`strata`" required by MAUSS, two options are given:

1. the frame, from which the sample will be selected, contains information about the target variables (the Y) survey (this is the case, for example, of frames containing census data or administrative data);

2. the frame does not contain such data: it will then need to calculate, for each stratum, the estimates for means and root mean square deviations of the Y’s, using different sources (for example, a previous round of the same survey, or different surveys with proxy estimates).

In the following, we examine both possibilities.

1. Availability of information concerning Y’s in the frame

In the R environment, a dataframe named “`frame`” contains the following information:

1. a unique identifier of the unit (no restriction on the name, may be “`cod`”);
2. the (optional) identifier of the stratum to which the unit belongs;
3. the values of m auxiliary variables (named from X1 to Xm);
4. the values of p target variables (named from Y1 to Yp);
5. the values of the domains of interest for which we want to produce estimates (named “`domainvalue`”).

For example:

```
> 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
```

If this information is available, it is possible to use the function `buildStrataDF` in this way:

```
> buildStrataDF(frame)
```

The function takes as argument the name of the single frame, and writes in the working directory the data frame containing information about strata (named "strata.txt"), structured as follows:

```
> 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. Availability of information from sources other than the frame (other surveys)

Conversely, if there is no information in the frame regarding the target variables, you must build the data frame "`strata`" from other sources, such as from a previous round of the same survey, or from other surveys.
In this case, assuming the information available is contained in a file named "samplePrev.txt", we need to read the data by running:

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

In addition to naming constraints introduced above, this feature requires that a variable named "`weight`" is present in the data frame "`samp`".
At this point you can perform the same function as already seen above:

```
> buildStrataDF(samp)
```

The result is much the same than the previous case: the function writes out in the working directory the strata file, named "strata.txt".

Note that in all cases, for each target variable Y, mean and standard deviation are calculated excluding NAs.