



Revision of

***“Guidelines for the protection of social microdata using the individual risk methodology”* by S. Poletini and G. Seri**

Silvia Poletini

ISTAT, MPS/D
Via C. Balbo, 16
00184, Roma
Italy

1. Introduction	2
2. Disclosure and disclosure scenario	3
3. Measures of risk.....	3
3.1. <i>Notation</i>	3
3.2. <i>Definition of disclosure risk for files of independent records</i>	4
3.3. <i>Estimation of the individual risk</i>	4
3.4. <i>Approximation of the individual risk.....</i>	6
3.5. <i>Assessing the risk of the whole file: global risk</i>	7
4. Application of local suppression within the individual risk methodology	8
4.1. <i>Threshold setting using the re-identification rate</i>	8
5. Releasing files with a hierarchical structure	9
5.1. <i>The household risk</i>	9
5.2. <i>Household risk: threshold setting and identification of the unsafe records</i>	10
6. Some Comments on Risk Estimation	11
7. A guided tour of μ-Argus for the application of the Individual Risk Methodology.....	11
7.1. <i>Specification of a household structure in the Menu Specify Metadata</i>	13
7.2. <i>Menu Specify Combinations.....</i>	15
7.2.1. <i>Output window</i>	17
7.3. <i>Menu Modify Risk Specification: the Risk Chart window.....</i>	19
7.3.1. <i>Risk histogram</i>	19
7.3.2. <i>Risk levels in the data before suppression.....</i>	19
7.3.3. <i>Setting the risk threshold to be used for local suppression.....</i>	19
7.4. <i>Menu Modify Household Risk Specification: the Household Risk Chart window.....</i>	23
7.4.1. <i>Setting the household risk threshold and identifying unsafe records</i>	23
7.4.2. <i>Validity of the assumptions that define the scenario for estimating the household risk.....</i>	24
7.5. <i>Output: Make protected file window.....</i>	24
References	27

1. Introduction

When microdata files are released, it is possible that external users may attempt to breach confidentiality. For this reason it is necessary to apply some form of disclosure risk assessment and data protection. In social surveys, the observed variables are frequently categorical in nature, and often comprise public domain variables (such as sex, age, region of residence) that may allow identification. An intruder may use these variables, that are referred to as *key variables*, to perform a *disclosure*. The definition of disclosure adopted in this framework uses the concept of *re-identification disclosure* (e.g. Chen and Keller-McNulty, 1998; Fienberg and Makov, 1998; Skinner and Holmes, 1998; Duncan and Lambert, 1986; Willenborg and de Waal, 2001): by disclosure it is meant a *correct record re-identification* operation that is achieved by an intruder when comparing a target individual in a sample with an available list of units that contains individual identifiers such as name and address. Re-identification occurs when the unit in the released file and a unit in the register that an intruder has access to belong to the same individual in the population. The underlying hypothesis is that the intruder will always try to match a record in the sample and a unit in the register using the *key variables* only.

This document describes the *individual risk methodology* as introduced in an initial paper by Benedetti and Franconi (1998) and implemented in version 4.0 of μ -Argus. This approach defines a measure of disclosure risk per-record (namely, the *individual risk*) that can be used to protect selected records by *local suppression*. The *individual risk* of re-identification of unit i in the sample is defined as the probability of it being correctly re-identified, i.e. recognised as corresponding to a particular unit in the population. In social surveys, this risk of re-identification can be expressed through the concept of *unique or rare combinations* in the sample. A combination is a cell in the contingency table obtained by cross-tabulating the key variables. A key issue is to be able to distinguish between combinations that are at risk, for example sample uniques corresponding to rare combinations in the population, and combinations that are not at risk, for example sample uniques corresponding to combinations that are common in the population. Benedetti and Franconi (1998) propose a framework for definition and estimation of the individual risk using the sampling weights, as the usual instrument that national statistical institutes adopt to allow for inference from the sample to the population

Besides the one just mentioned, a number of proposals for defining and estimating a re-identification risk per record has been made in the last few years: Fienberg and Makov (1998), Skinner and Holmes (1998), Elamir and Skinner (2004) define, with different motivations, a log linear model for the estimation of the individual risk. Further discussion of the approach presented here is in Rinott (2003), Polettini (2003b), Franconi and Polettini (2004). A related approach is described in Carlson (2002), Elamir and Skinner (2004) and Polettini and Stander (2004).

After the risk has been estimated, protection takes place. To this aim, a threshold in terms of risk, e.g. probability of re-identification (see Section 3.2) is selected; units exceeding such a threshold are defined at risk, and μ -Argus applies local suppression to those individuals only, so as to lower their probability of being re-identified. Note that this approach allows to release the sensitive variables unchanged, while suppressing some of the key variables for some records.

The paper has two different components: a technical description of the individual risk methodology, and a guided tour through μ -Argus for the application of the individual risk methodology

The first part is devoted to the methodology. Section 2 introduces the definition of disclosure and the disclosure scenario. Section 3 is devoted to measures of risk for independent records: after introducing the basic notation (Section 3.1), the individual risk is presented in Sections 3.2 (definition), 3.3 (estimation) and 3.4 (approximation). Section 3.5 introduces a new concept of global risk of the whole file, namely the re-identification rate, that can be helpful in selecting a threshold for the individual risk; this topic is dealt with in Section 4. Section 5 discusses the release

of household data: the household risk is presented in Section 5.1, while Section 5.2 indicates how to set the threshold for household data. Finally, Section 6 contains a discussion of some theoretical issues in the individual risk methodology.

The second part is devoted to application and guides the reader through the use of μ -Argus; in particular, Section 7 shows the main steps needed to produce a safe microdata file according to the individual risk methodology, that combines local suppression with the individual risk measure .

2. Disclosure and disclosure scenario

As mentioned in the Introduction section, the definition of *disclosure* mimics the strategy of an intruder trying to establish a link between a unit in the sample s to be released and a unit in an available archive. Such an archive, or *register*, contains individual direct identifiers (*name, ID number, phone number...*) plus a set of variables called *identifying* or *key variables* (*sex, age, marital status...*). The intruder tries to match unit i in the sample with a unit i^* in the register by comparing their scores on the key variables. A *re-identification* occurs when, based on this comparison, a unit i^* in the register is selected as matching to i and *this link is correct*, e.g. i^* is the labelling of unit i in the population.

To define the *disclosure scenario*, the following assumptions are made. Most of them are conservative and contribute to the definition of a worst case scenario:

- 1) a *sample* s from a population P is to be released, and *sampling design weights* are available;
- 2) the archive available to the intruder covers *the whole population* P ; consequently for each $i \in s$, the matching unit i^* does always exist in P ;
- 3) the archive available to the intruder contains the individual direct identifiers and a set of categorical *key variables* that are also present in the sample;
- 4) the intruder tries to match a unit i in the sample with a unit i^* in the population register by comparing the values of the key variables in the two files;
- 5) the intruder has no extra information other than that contained in the register;
- 6) a *re-identification* occurs when a link between a sample unit i and a population unit i^* is established and i^* is actually the individual of the population from which the sampled unit i was derived; e.g. the match has to be a *correct match* before an identification takes place.

Moreover we add the following assumptions:

- 7) the intruder tries to match all the records in the sample with a record in the population register;
- 8) the key variables agree on correct matches, that is no errors, missing values or time-changes occur in recording the key variables in the two data archives.

3. Measures of risk

In this section we briefly introduce the basic concepts underlying the individual risk methodology within μ -Argus. For a detailed description of the individual risk methodology we refer to Benedetti and Franconi (1998), Benedetti, Franconi and Capobianchi (2003), Franconi and Poletini (2004). The concepts of individual and global risk are introduced, and technical details are provided.

3.1. Notation

Let the released file be a random sample s of size n selected from a finite population P consisting of N units. For a generic unit i in the population, we denote by $1/w_i$ its probability to be included in the sample.

Consider the contingency table built by cross-tabulating the key variables. A *combination* k is defined as the k -th cell in the contingency table. The set of combinations $\{1, \dots, k, \dots, K\}$ defines a *partition* of both the population and the sample into cells. Observing the values of the key variables on individual $i \in s$ will classify such individual into one cell. We denote by $k(i)$ the index of the cell

into which individual $i \in s$ is classified based on the values of the key variables. Typically, we expect to find several sampled units within the same combination k . We focus the analysis on each of the $k=1, \dots, K$ cells of this contingency table.

Let f_k and F_k denote, respectively, the number of records in the released file and the number of units in the population with the k -th combination of categories of the key variables; F_k is unknown for each k . Depending on the key variables, the total number K of combinations can be quite high; in the sample to be released, only a subset of such number will be observed and only this subset of combinations for whom $f_k > 0$, is of interest to the disclosure risk estimation problem.

3.2. Definition of disclosure risk for files of independent records

According to the concept of re-identification disclosure mentioned, we define the individual risk of disclosure of unit i in the sample as its *probability of re-identification*. In symbols:

$$\rho_i = \Pr(i \text{ correctly linked with } i^* \mid s, P). \quad (1)$$

Clearly the probability that $i \in s$ is correctly linked with $i^* \in P$ is null if the intruder does not perform any link. Therefore we can condition on the event L_i : “the intruder attempts a re-identification of unit $i \in s$ ” and write

$$\rho_i = \Pr(i \text{ correctly linked with } i^* \mid s, P, L_i) \Pr(L_i),$$

where $\Pr(L_i)$ represents the probability that the intruder tries to establish a link between unit $i \in s$ and some unit in P .

The re-identification attempt is cast under the scenario described in Section 2. We adopt a pessimistic *scenario* by assuming that the intruder attempts to match all the records in the file ($\Pr(L_i)=1$ for all i); moreover we assume that the key variables agree on matching pairs, e.g. these are recorded without error or missing values in either s or P and moreover no time changes occur in the values of the key variables. The latter hypothesis raises the probability of re-identification of record i . Therefore the risk r_i that we get under this scenario is certainly not smaller than the risk ρ_i of formula (1):

$$\rho_i \leq r_i = \Pr(i \text{ correctly linked with } i^* \mid s, P, \text{worst case scenario}) \quad (2)$$

Therefore measuring the disclosure risk by r_i in place of ρ_i is *prudential*, e.g. the actual risk is lower than the one we are estimating. We refer to r_i in (2) as the (base) *individual risk of re-identification*; this is the measure available in μ -Argus. Recall that r_i is an upper bound to the probability of re-identification of unit i in the sample. Details on estimation of r_i are provided in the next section.

3.3. Estimation of the individual risk

We take into account the base individual risk of formula (2), that is, the upper bound to the probability of re-identification of a unit in the sample under the worst case scenario described in Section 2. As we already discussed, the risk is cast in terms of the cells of the contingency table built by cross-tabulating the key variables. Consequently all the records in the same cell have the same value of the risk; for this reason, for any record i in cell k we refer to r_k instead of r_i .

Looking at the k -th combination of the key variables, $k=1, \dots, K$, the intruder finds in the sample f_k individuals having such combination, out of F_k in the population. These individuals are exchangeable for re-identification, e.g. each of the F_k can be linked to any of the f_k . If we were to know the population frequency of the k -th combination, F_k , we would define the probability of re-identification simply by $1/F_k$. The agency that distributes the data may not have access to the population register, and may not know the population cell sizes F_k , therefore an inferential step is to be performed. We define the individual risk as *the agency estimate of the upper bound to the intruder's re-identification probability*. In the proposal by Benedetti and Franconi (1998), the uncertainty on F_k is accounted for in a Bayesian fashion by introducing the distribution of the population frequencies given the sample frequencies. The individual risk of disclosure is then

measured as the (posterior) mean of $1/F_k$ with respect to the distribution of $F_k|f_k$:

$$r_i = E\left(\frac{1}{F_k} \mid f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \Pr(F_k = h \mid f_k). \quad (3)$$

To determine the probability mass function of $F_k|f_k$, the following superpopulation approach is introduced (see Bethlehem *et al.*, 1990; Rinott, 2003; Poletini, 2003b):

$$\begin{aligned} \pi_k \square [\pi_k] &\propto 1/\pi_k && \text{independently, } k = 1, \dots, K \\ F_k \mid \pi_k &\square \text{Poisson}(N\pi_k) && \text{independently, } F_k = 0, 1, \dots \\ f_k \mid F_k &\square \text{binomial}(F_k, p_k) && \text{independently, } f_k = 0, 1, \dots, F_k. \end{aligned} \quad (4)$$

Under model (4), the posterior distribution of $F_k|f_k$ is negative binomial with success probability p_k and number of successes f_k . The law of $F_k|f_k$ is that of a negative binomial variable counting the number of trials before the j -th success, each with probability p_k . Its probability mass function is:

$$\Pr[F_k = h \mid f_k = j] = \binom{h-1}{j-1} p_k^j (1-p_k)^{h-j}, \quad h \geq j. \quad (5)$$

In Benedetti and Franconi (1998) it is shown that under the negative binomial distribution (5) the risk (3) can be expressed as

$$r_k = E(F_k^{-1} \mid f_k) = \int_0^{\infty} \left\{ \frac{p_k e^{-t}}{1 - q_k e^{-t}} \right\}^{f_k} dt \quad (6)$$

where $q_k = 1 - p_k$.

In the original formulation, the transformation $y = (1 - q_k e^{-t})^{-1}$ and the Binomial theorem were used in (6) to get

$$r_k = \left(\frac{p_k}{q_k} \right)^{f_k} \int_1^{1/p_k} \frac{1}{y} (y-1)^{f_k-1} dy = \left(\frac{p_k}{q_k} \right)^{f_k} \left\{ \sum_{j=0}^{f_k-2} (-1)^j \binom{f_k-1}{j} \frac{p_k^{j+1-f_k} - 1}{f_k - j - 1} + (-1)^{f_k} \log(p_k) \right\} \quad (7)$$

which is valid for $f_k > 1$.

Alternatively, formula (6) can be expressed via the transformation $y = \exp(-t)$ as:

$$r_k = p_k^{f_k} \int_0^1 t^{f_k-1} (1 - tq_k)^{-f_k} dt; \quad (8)$$

further, (8) can be rewritten (Poletini, 2003a) in terms of the hypergeometric function as

$$r_k = \frac{p_k^{f_k}}{f_k} {}_2F_1(f_k, f_k; f_k + 1; q_k) \quad (9)$$

where

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1} (1 - tq_k)^{-a} dt$$

is the integral representation (valid for $\Re(c) > \Re(b) > 0$) of the Gauss hypergeometric series (see Abramowitz and Stegun, 1965).

An estimate of the individual risk of disclosure (3) can be obtained by estimating p_k in (7) or (9). Given F_k , the maximum likelihood estimator of p_k under the binomial model in (4) is

$$\hat{p}_k = \frac{f_k}{F_k}.$$

F_k being not observable, Benedetti and Franconi (1998) propose to use

$$\hat{p}_k = \frac{f_k}{\sum_{i:k(i)=k} w_i}, \quad (10)$$

where $\sum_{i:k(i)=k} w_i$ is an estimate of F_k based on the sampling design, possibly calibrated (Deville and

Särndal, 1992).

Generally speaking, use of formula (7) for the risk leads to numerically unstable estimates for values of \hat{p}_k close to 0 or 1. Formula (9) does not suffer from this drawback. Using (10) in (9) we therefore end up with the following *plug-in estimate of the individual risk*:

$$\hat{r}_k = \frac{\hat{p}_k^{f_k}}{f_k} {}_2F_1(f_k; f_k; f_k + 1; 1 - \hat{p}_k). \quad (11)$$

The negative binomial distribution is defined for $0 < p_k < 1$; in practice, the estimates \hat{p}_k might attain the extremes of the unit interval. We never deal with $\hat{p}_k = 0$, as it corresponds to $f_k = 0$; on the other hand, if $\hat{p}_k = 1$, ${}_2F_1(f_k; f_k; f_k + 1; 1 - \hat{p}_k) = 1$, so that the individual risk equals $1/f_k$.

3.4. Approximation of the individual risk

For large values of the parameters f_k , $1 - \hat{p}_k$, numerical evaluation of the hypergeometric function can be computationally demanding. Poletti (2003b) has derived approximations that can be used even for moderate cell sizes. The approximations provided are based on the series representation of the hypergeometric function ${}_2F_1(f_k, f_k; f_k + 1; q_k)$. The former is divergent when $f_k < 0$, therefore divergence is never of concern in practice. Absolute convergence of the series is guaranteed for $f_k > 1$. This constraint does never apply as we are dealing with an approximation; moreover the analytic expression for the risk when $f_k = 1$ is known and equals $-\log(p_k) \frac{p_k}{1 - p_k}$.

For $f_k = 2$ or 3 the analytic expressions of the risk are known as well:

$$f_k=2: \quad r_k = \frac{p_k}{q_k^2} (p_k \log p_k + q_k); \quad f_k=3: \quad r_k = \frac{p_k}{2q_k^3} (q_k(3q_k - 2) - 2p_k^2 \log p_k).$$

The approximations to be proposed use the contiguity property of the hypergeometric function

$${}_2F_1(f_k, f_k; f_k + 1; q_k) = (1 - q_k)^{1-f_k} {}_2F_1(1, 1; f_k + 1; q_k)$$

(see Abramowitz and Stegun, 1965), and the series expansion:

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^n}{n!}.$$

The formulae above lead to expressions of the type

$$r_k = \frac{p_k}{f_k} \left(1 + \frac{q_k}{f_k + 1} + \frac{2q_k^2}{(f_k + 1)(f_k + 2)} + \frac{6q_k^3}{(f_k + 1)(f_k + 2)(f_k + 3)} + O(f_k^{-4}) \right). \quad (12)$$

Approximations from below can be derived by truncating the series representation (12); the error depends on the remainder. Better accuracy may be achieved by introducing additional terms in the truncated series representation. In general, the approximation

$$\frac{p_k}{f_k} \left(1 + \frac{q_k}{f_k + 1} + \frac{2q_k^2}{(f_k + 1)(f_k + 2)} \right) \quad (13)$$

which has order $O(f_k^{-3})$, is accurate even for moderate cell sizes. Numerical assessment of the relative error $(r_k - r'_k)/r_k$ has shown that the fourth order representation

$$r'_k = \frac{p_k}{f_k} \left(1 + \frac{q_k}{f_k + 1} + \frac{2q_k^2}{(f_k + 1)(f_k + 2)} + \dots + \frac{7!q_k^7}{(f_k + 1)(f_k + 2) \cdots (f_k + 7)} \right) \quad (14)$$

is satisfactory even for small cell sizes (see Fig. 1). Figure 1 also indicates that the truncated series approximation is not uniform in f_k and p_k and in general is more accurate for high values of f_k and small values of q_k .

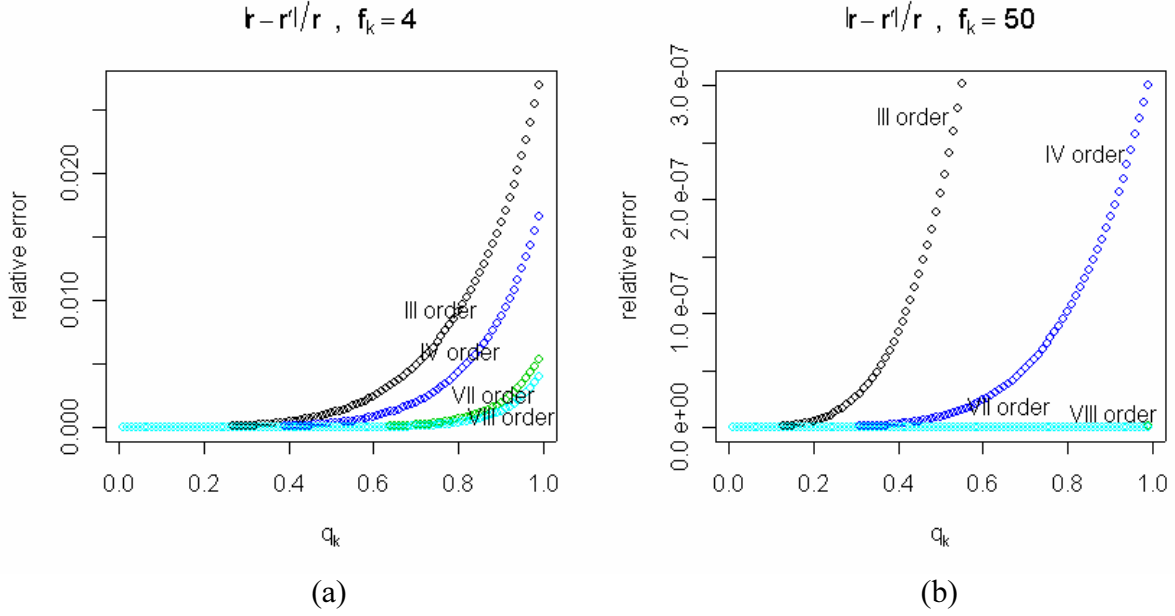


Figure 1. Relative errors of the approximation for different orders (a) $f_k = 4$ and (b) $f_k = 50$

Evaluation of the risk using the hypergeometric function requires specialised code and can prove computationally intensive for large values of the parameters; for these reasons and based on the previous findings, we suggest using formula (14) to evaluate the risk. Estimation of the risk can be performed by plug-in of the estimates \hat{p}_k of formula (10) in the approximation (14).

3.5. Assessing the risk of the whole file: global risk

The individual risk provides a measure of risk *at the individual level*. A *global* measure of disclosure risk for the whole file can be expressed in terms of the expected number of re-identifications in the file (see Lambert, 1993). In this section we introduce the *expected number of re-identifications* and the *re-identification rate*. Whereas the first is a measure of disclosure that depends on the number of records in the file, the re-identification rate is independent of n .

Define a dichotomous random variable Φ , assuming value 1 if the re-identification is correct and 0 if the re-identification is not correct. In general for each unit in the sample one such variable Φ_i is defined, assuming value 1 with at most probability r_i . In the discussion we behave as if such probability was *exactly* r_i . The random variables Φ_i are not *i.i.d.*, but the risk is constant over the cells of the contingency table; therefore for combination k of the key variables we have f_k *i.i.d.* random variables Φ_i assuming value 1 when the re-identification is correct, with constant probability r_k . This probability can be exploited to derive the *expected number of re-identifications per cell*, which equals $f_k r_k$. In general, the overall expected number of re-identifications over the whole sample is

$$ER = \sum_{i=1}^n E(\Phi_i) = \sum_{k=1}^K f_k r_k.$$

The previous measure can be used to define the *re-identification rate* ξ as

$$\xi = \frac{1}{n} ER = \frac{1}{n} \sum_{k=1}^K f_k r_k. \quad (15)$$

ξ provides a measure of *global risk*, *i.e.* a measure of disclosure risk for the whole file, that does not depend on the sample size and can be used to assess the risk of the file or to compare different types of release.

The *percentage of expected re-identifications*, *i.e.* the value $\psi = 100 * \xi$ % provides an equivalent

measure of global risk.

4. Application of local suppression within the individual risk methodology

μ -Argus contains a module that estimates the individual risk. As we have already seen, actually what is estimated is an upper bound for the probability of re-identification, thus leading to *prudential evaluation* of the risk. After the risk has been estimated, protection takes place. One option in protection is recoding the categories of selected variables (*global recoding*), the other is the application of *local suppression*. Local suppression consists of introducing missing values in some of the key variables of selected records; note that this choice allows to release the sensitive variables unchanged.

In μ -Argus the user can combine the technique of local suppression with either of two different measures of risk, both at the combination level: one is the sample frequency of the combination (used by the CBS methodology), the other is the risk of the individuals in that combination. In either event, protection by local suppression is applied to unsafe cells or combinations, and the user must input a *threshold* to classify these as either safe or unsafe. When applying local suppression in combination with the individual risk, users must select a threshold in terms of risk, e.g. probability of re-identification (see Section 3.2). Local suppression is applied to the unsafe individuals, so as to lower their probability of being re-identified. In order to select the risk threshold, that represents a level of *acceptable risk*, i.e. a risk value under which an individual can be considered safe, we suggest that the user refers to the *re-identification rate* defined in the previous section. Indeed the re-identification rate has a very natural interpretation, and a fixed range, whereas the range of the individual risk varies substantially with the data, and it is usually not possible to define a threshold that is valid for all data releases.

The user will therefore define a *release safe* when the expected rate of correct re-identifications is below a level he/she considers acceptable. As the re-identification rate is cast in terms of the individual risk (see formula (15)), a threshold on the re-identification rate can be transformed into a threshold on the individual risk: this is described in Section 4.1. Under this approach, individuals are at risk because their probability of re-identification contributes a large proportion of expected re-identifications in the file.

Clearly, choice of the threshold affects the quality of the resulting “safe” file, and before saving the output file, the threshold can be changed. This will allow for assessment of the risk of the file, number of consequent suppressions and therefore quality of the “safe” file before data release. In order to reduce the number of suppressions, joint use of *global recoding* and local suppression is recommended. Recoding of selected variables will indeed lower the individual risks and therefore the re-identification rate of the file. The whole procedure of releasing a safe file indeed makes use of both global recoding and local suppression (see Section 7). The individual risk has here a twofold use: directly, it permits to identify unsafe records; indirectly, as r_i measures the contribution of record i to the re-identification rate ξ (or its equivalent ψ), it permits to assess whether the whole file can be safely released.

4.1. Threshold setting using the re-identification rate

In the previous section we remarked that a threshold for the individual risk can be determined by choosing the maximum tolerable *re-identification rate* in the sample. We next motivate this correspondence and show how this operation can be performed.

Consider the re-identification rate (15): cell k contributes to ξ an amount $r_k f_k$ of expected re-identifications. Since units belonging to the same cell k have the same individual risk, we can arrange cells in increasing order of risk r_k . We use the brackets to denote the generic element in this ordering; therefore the k -th element in this ordering will be denoted by the subscript (k) .

Suppose that a *threshold* r^* has been set *on the individual risk*. Unsafe cells are those for which $r_k \geq r^*$. In the risk ordering of cells, we can find a cell index (K^*) (which is relative to r^*) that

discriminates between safe and unsafe cells, i.e.

$$K^* \text{ such that } r_{(K^*)} < r^* \text{ and } r_{(K^*+1)} \geq r^* ; \quad (16)$$

Formula (16) above puts r^* and K^* in one-to-one correspondence. Unsafe cells are indexed by $(k) = K^*+1, \dots, K$, and local suppression ensures that after protection these cells will have risk below r^* . For the unsafe cells therefore $r_{(k)} < r^*$ *after protection*. Once protection is applied using threshold r^* , the expected number of re-identifications in the *released* sample is then certainly smaller than

$$\sum_{(k)=1}^{K^*} f_{(k)} r_{(k)} + r^* \sum_{(k)=K^*+1}^K f_{(k)} .$$

In fact, the threshold r^* can be picked from the discrete set of *observed* risk values, because choosing a threshold that is bracketed by two consecutive observed risk levels would not change the expected number of re-identifications.

The previous formula allows setting r^* so that the re-identification rate of the released file is bounded by a user-defined threshold value, i.e. $\zeta < \zeta^*$. A sufficient condition for this is that

$$\frac{1}{n} \left(\sum_{(k)=1}^{K^*} f_{(k)} r_{(k)} + r^* \sum_{(k)=K^*+1}^K f_{(k)} \right) < \zeta^* . \quad (17)$$

Instead of selecting a threshold on the individual risk, it is more natural for the user to define a *threshold on the re-identification rate* ζ . Reversing the approach pursued so far, formula (17) can be exploited to determine a cell index K^* that keeps the re-identification rate ζ of the released file below τ . This in turn identifies a threshold on the individual risk: by relation (16), the corresponding individual risk threshold is $r_{(K^*+1)}$. The search of such a K^* is performed by a simple iterative algorithm.

5. Releasing files with a hierarchical structure

A relevant characteristic of social microdata is its inherent hierarchical structure, which allows us to recognise groups of individuals in the file, the most typical case being the *household*. Very often in social surveys the same information is collected for each household member and all this is stored in a single record that refers to the household. When defining the re-identification risk, it is important to take into account this dependence among units: indeed re-identification of an individual in the group may affect the probability of disclosure of all its members. So far, implementation of a hierarchical risk has been performed only with reference to households. We will therefore refer to *individual* and *household risk* for files of independent units and of households, respectively.

The individual risk methodology is currently the only approach that to some extent allows for a hierarchical structure in the data, which is in fact typical in many contexts. Under the hypothesis that the file has a hierarchical structure, it is possible to locate units within the household and, to a certain extent, also establish relationships between members of the same household. Allowing for dependence in estimating the risk enables us to attain a higher level of safety than when merely considering the case of independence. In the next section we introduce a measure that addresses this problem, in the effort to suit hierarchically structured, in particular household, microdata.

5.1. The household risk

For the household risk we adopt the same framework that we referred to in defining the individual risk. In particular, we use the concept of re-identification disclosure and the scenario described in Section 2. However, as external registers that contain information on households are not commonly available, for files of households we make the additional assumption that

- 9) the intruder attempts a confidentiality breach by re-identification of *individuals* in households.

From the definition given in Section 3.2, it follows that the individual risk can be considered an estimate of a probability of re-identification. We define the *household risk* as the probability that *at least* one individual in the household is re-identified. By consequence, the household risk can be derived from the individual risks and knowledge of the household structure of the data. For a given household g of size $|g|$, whose members we label $i_1, \dots, i_{|g|}$, we define the household risk as

$$r^h(g) = Pr(i_1 \cup i_2 \cup \dots \cup i_{|g|} \text{ re-identified}). \quad (18)$$

Assuming independence of re-identification attempts within the same household, (18) can be expressed by Boole's formula using the individual risks $r_{i_1}, \dots, r_{i_{|g|}}$ defined in (3):

$$r_g^h = \sum_{j=1}^{|g|} r_{i_j} - \sum_{i_j < i_i} r_{i_j} r_{i_i} + \sum_{i_j < i_i < i_m} r_{i_j} r_{i_i} r_{i_m} + \dots + (-1)^{|g|+1} r_{i_1} r_{i_2} \dots r_{i_{|g|}} \quad (19)$$

By symmetry of the Boole's formula, the ordering of units in the group is not relevant. In a hierarchical file therefore the measure of disclosure risk is the same for all the individuals in household g and equals r_g^h . Estimation of the household risk can be performed by plug-in of the estimates of the individual risks $r_{i_1}, \dots, r_{i_{|g|}}$ along the lines described in Section 3.

5.2. Household risk: threshold setting and identification of the unsafe records

Since all the individuals in a given household have the same household risk, the expected number of re-identified records in household g equals $|g| r_g^h$. We define the re-identification rate in a hierarchical file as

$$\xi^h = \frac{1}{n} \sum_{g=1}^G |g| r_g^h,$$

where G is the total number of households in the file. The re-identification rate can now be used to define a threshold r^{h*} on the household risk r^h , much in the same way as in Section 4.1.

Note that the household risk r_g^h of household g is computed by the individual risks of its household members. For a given household, it might happen that a household is unsafe (r_g^h exceeds the threshold) because just one of its members, i , say, has a high value r_i of the individual risk. In order to protect the households, our approach is therefore to protect individuals in households, first protecting those individuals who contribute most to the household risk. For this reason, inside *unsafe households*, we need to identify *unsafe individuals*. In other words, we need a way to transform a threshold on the household risk r^h into a threshold on the individual risk r . To this aim, we notice that by formula (19) the household risk is bounded by the sum of the individual risks of the members of the household:

$$r_g^h \leq \sum_{j=1}^{|g|} r_{i_j}.$$

Consider to apply a threshold r^{h*} on the household risk. In order for household g to be classified safe (i.e. $r_g^h < r^{h*}$) it is *sufficient* that all of its components have individual risk less than

$$\delta_g = r^{h*} / |g|.$$

This is clearly a strongly prudential approach, as we check whether a *bound* on the household risk is below a given threshold.

It is important to remark that the threshold δ_g just defined depends on the size of the household to which individual i belongs. This implies that for two individuals that are classified in the same combination k of key variables (and therefore have the same individual risk r_k), but belong to different households with different sizes, it might happen that one is classified safe, while the other unsafe.

In practice, denoting by $g(i)$ the household to which record i belongs, the approach pursued so far consists in turning a threshold r^h on the household risk into a *vector of thresholds* on the *individual risks* r_i ; $i = 1, \dots, n$:

$$\delta_g = \delta_{g(i)} = r^h / |g(i)|.$$

Individuals are finally set to unsafe whenever $r_i \geq \delta_{g(i)}$; local suppression is then applied to those records, if requested. Suppression of these records ensures that after protection the household risk is below the threshold δ_g .

6. Some Comments on Risk Estimation

The procedure relies on the assumption that the available data are a sample from a larger population. The sampling design is assumed to be known, as far as the sampling weights are concerned at least. *If the sampling weights are not available, or if data represent the whole population, the strategy used to estimate the individual risk is not meaningful.* Therefore we recommend using the individual risk methodology only for *sample data, when sampling weights are available.*

Recall that the household risk r_g^h of household g only depends on the risks $r_{i_1}, \dots, r_{i_{|g|}}$ of its components, and these are evaluated under the hypothesis of no hierarchical structure in the file. For household data it is therefore important to include in the key variables that are used to estimate the risks $r_{i_1}, \dots, r_{i_{|g|}}$ also the available information on the household, such as the number of components or the household type. Note that when the household size is used as one of the key variables, a single threshold δ_g for each cell is defined, as all the individuals in the same cell have the same household size. Under this circumstance records in cell k are all safe or all unsafe (contrast with the remark of Section 5.2).

Suppose one computes the risk using the household size as the only key variable in a household data file, and that such file contains households whose risk is above a fixed threshold. Since information on the number of components in the household cannot be removed from a file with household structure, these records cannot be safely released, and no suppression can make them safe. This permits to check for presence of very peculiar households (usually, the very large ones) that can be easily recognised in the population just by their size and whose main characteristic, namely their size, can be immediately computed from the file.

In Di Consiglio et al. (2003) an experiment was conducted to assess the performance of the individual risk of disclosure. The aim was to investigate whether the individual risk is estimating the correct quantity, i.e. the real risk of an individual, and also whether the quality of this estimation is appropriate. A good agreement was noticed between the real risk and its estimates, although the precision of the estimator seems poor for rare combinations in the sample as compared to the more common ones. A minor precision is an intrinsic problem of small counts. Discriminating rare and common features in the population is, by far, the most difficult task especially when one can count on only one occurrence in the sample. For this reason further studies to improve the performance of the estimator used for the individual risk have been planned. Model (4) was questioned to provide a good fit to real data under some circumstances (e.g. Rinott, 2003). Our experiments seem to indicate that the model holds, at least with the data we use. Notice that the assumed model is compatible with a large number of relatively small cells in the population. These might occur with both small sized population and large number of combinations of key variables.

7. A guided tour of μ -Argus for the application of the Individual Risk Methodology

This Section describes application of the individual risk methodology within μ -Argus version 4.0. Basically, the method detailed in the previous sections defines a measure of re-identification risk per record that is estimated using the sampling design weights. We distinguish here two different

types of application: the first concerns files of *independent records*, whereas the second concerns files with a *household structure*. The latter must have a special structure that we discuss in Section 7.1. A household structure is detected by the software through the presence of a **household identifier** variable, i.e. a counter that distinguishes different households. This information has to be imputed in the metadata window (see Section 7.1). A household file usually also presents **household variables**, i.e. variables that take the same level for any member of the household.

For files of independent records, μ -Argus computes the individual risk measure described in Section 3; for household data, μ -Argus automatically computes a household risk as described in Section 5.

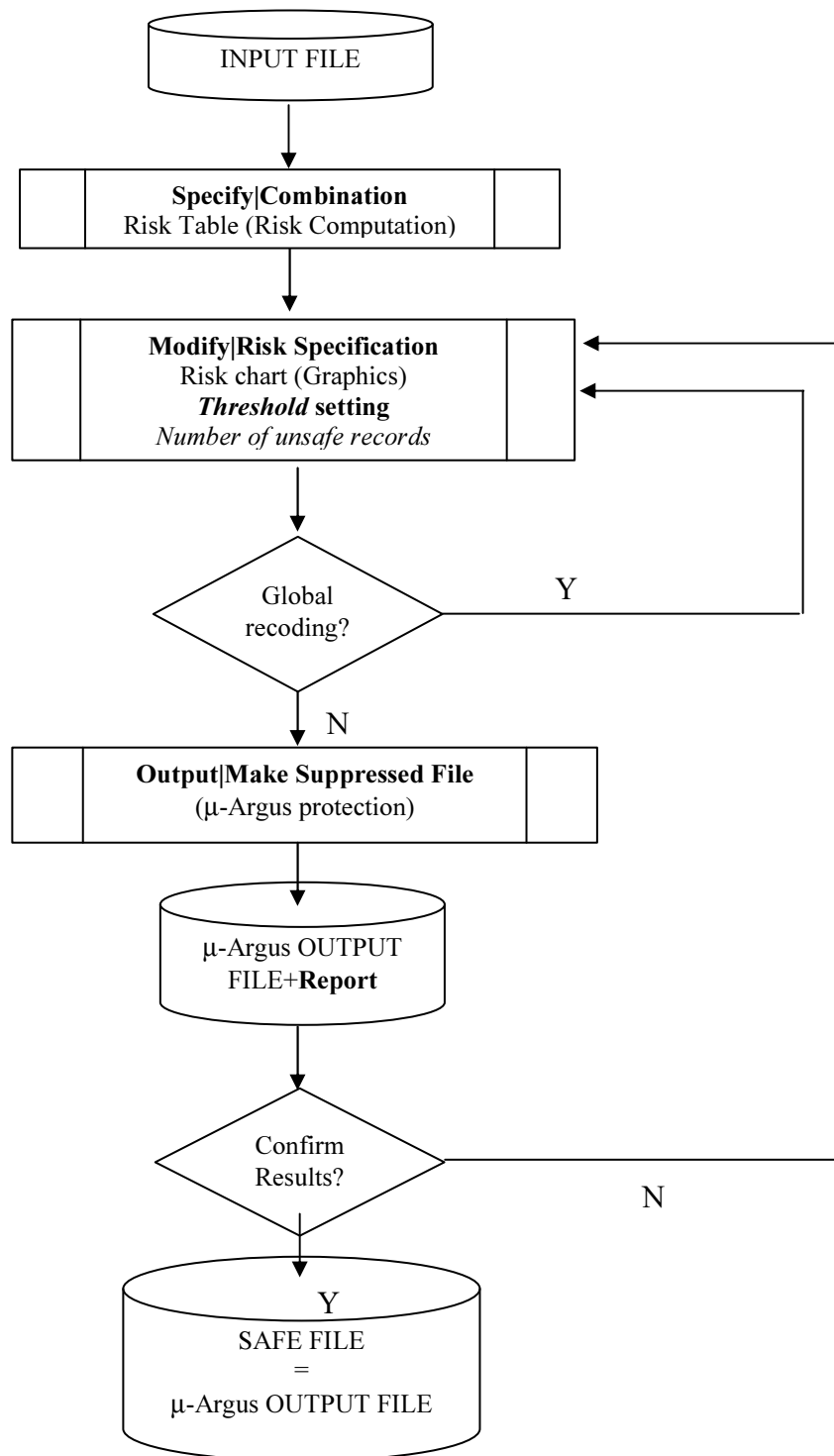
Once the proper risk measure has been estimated, *local suppression* is applied to the individuals whose risk exceeds a given threshold. This is to be set by the user (see Sections 4 and 5). Local suppression lowers the individual risk of the protected records, and therefore the global risk of the file, as measured by the re-identification rate. To reduce the number of suppressions, *global recoding* can also be applied. Global recoding affects all the records in the file, thereby reducing their risks; as a consequence, the re-identification rate or global risk of the file is decreased.

We next discuss the steps needed to perform the protection of a data file under the individual risk model. Generally speaking, μ -Argus requires the following elements to be specified:

- the data and its structure;
- the *key variables*;
- a *risk table* based on the observed combinations of the key variables, that contains the individual risks. As the individual risk is the same for all the individuals having the same combination of key variables, the risk table can be built on combinations;
- a *threshold* that permits to classify records into safe/unsafe, respectively.

Recall that the *key variables* are public domain variables, such as sex, age, region of residence, that may allow identification; we assume that these are available to the intruder when attempting a disclosure.

The next Sections discuss the main ingredients of the procedure, presented following the steps that a user of μ -Argus has to take in order to protect the data using the individual risk method together with local suppression and, possibly, global recoding. The graph shows the main lines along which the individual risk methodology proceeds.



7.1. Specification of a household structure in the Menu Specify|Metadata

The procedure for importing files of independent records does not differ from what is described in Section 4 of the user manual of μ -Argus. The menu **Specify|Metadata** is indeed common to any protection procedure applied by the software. For this reason we only discuss the case when the file to be protected has a *household structure*.

First of all, the data file must be presented in the form of a *file of individuals*, in which an additional

variable, namely the **household identifier**, contains the household structure of the data. For an example of one such file, see the scheme in Figure 2. A few lines of the SAMHH.ASC demonstration file (that can be found in the hhdata directory of the μ -Argus installation directory) are also shown in Figure 3 to see how the input file looks like. The household identifier (highlighted in both figures) is simply a counter that permits to identify the household and to distinguish different households.

hhident	region	sex	marstat	age	hhtype
1	1	2	1	43	2
2	7	1	3	40	2
3	3	1	1	47	5
3	3	2	1	48	5
4	5	2	1	63	11
4	5	1	1	61	11
4	5	1	1	57	11
4	5	1	1	48	11
4	5	2	1	37	11
5	5	1	2	61	8
5	5	2	2	64	8
5	5	2	1	32	8
5	5	2	1	30	8
6	8	2	6	57	10
6	8	2	1	32	10
6	8	1	4	38	2
7	11	1	6	77	3

Figure 2. Structure of a household data file. **hhident** is the household identifier whereas **region** and **hhtype** represent the household variables “region of residence” and “household type”.

```

122.2221 11152 0 40
122.2221 32401 150 40
122.2221 12112 0 40
266.6662 22762 0 76
266.6662 21712 0 76
355.5552 213911588 39
355.5552 22402 0 39
355.5552 11132 0 39
355.5552 12112 0 39
466.6662 11221 878 22
511.1111 21261 493 26
511.1111 22271 168 26
511.1111 11 22 0 26
511.1111 12 22 0 26
511.1111 12 12 0 26
677.7772 213611477 36
677.7772 22362 0 36
677.7772 12 92 0 36
677.7772 12 82 0 36
744.4442 114112729 41
855.5552 21241 979 24
855.5552 22222 0 24
855.5552 12 02 0 24

```

Figure 3 Structure of the **hhdata**. Highlighted is the household identifier variable.

It is important to remark that **the input file must be arranged into households**, in the sense that it is necessary that **at least the members of the same household are grouped together in the input**

file. A sufficient condition for that is to sort the file by the value of the household identifier variable. If this is not the case, the household risk that Argus computes will be unreliable. The information on the household identifier, which is crucial to make Argus aware that a household structure is present in the data, has to be entered in the **Specify|Metadata** window (setting for this variable the type **HH identifier**, see Figure 4). A household data file generally also presents **household variables** (variables that take the same level for all the members of the household, see Figure 2 and Section 2.7 of the manual); this information can be specified in the same window. The other steps do not differ from the usual data import procedure.

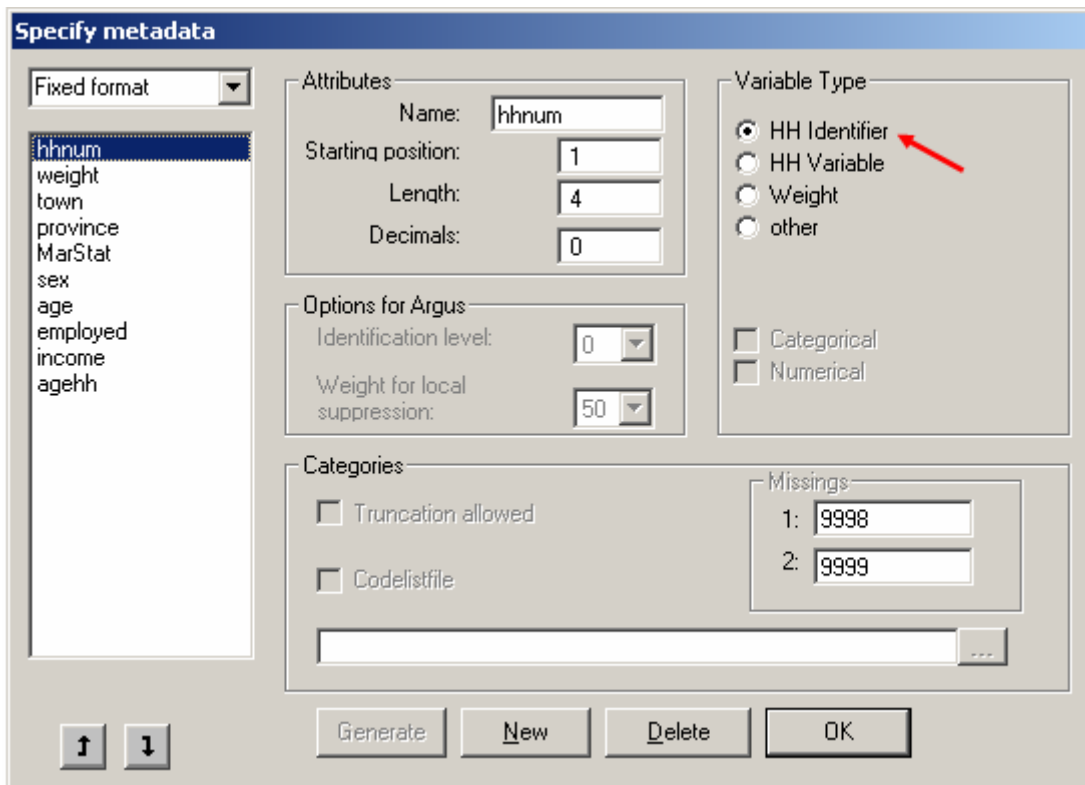


Figure 4. Specifying the household identifier variable **hhnum** to let μ -Argus read the household structure of the file

7.2. Menu Specify|Combinations

From the menu **Specify|Combinations**, users access a multi-purpose window (**Select Combinations**, see Figure 5) where the key variables can be selected. These represent the information that is assumed to be available to the intruder and therefore depend on the scenario. The software completes this step by building the contingency table obtained by cross-classifying the selected key variables; if the user requires to apply the *individual risk methodology*, μ -Argus additionally stores a so-called **risk table**, i.e. a contingency table in which a risk value is associated to each cell or combination. Note that μ -Argus allows specifying separate subgroups of key variables; in this circumstance, the software computes as many contingency tables as there are groups of key variables. When the traditional (Dutch) method is chosen, μ -Argus computes the unsafe combinations according to the concept of *uniques* or *fingerprints* as described e.g. in Willenborg and de Waal (1996; 2001). Under the individual risk approach, the unsafe records are those exceeding a given risk threshold; this can only be specified after the risk has been computed (see Section 7.3)

From the left panel of the window, the key variables can be moved to the central box using the button **1** in Figure 5. Move further the selected set of key variables to the right panel, using the

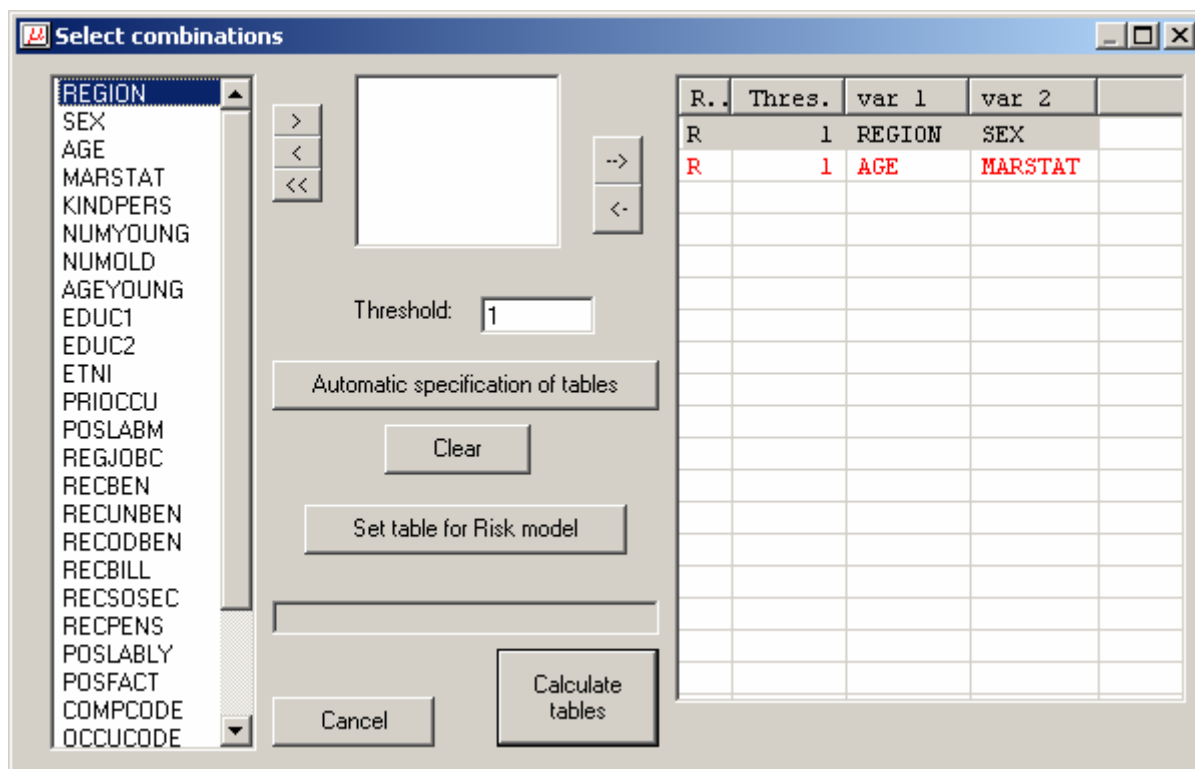
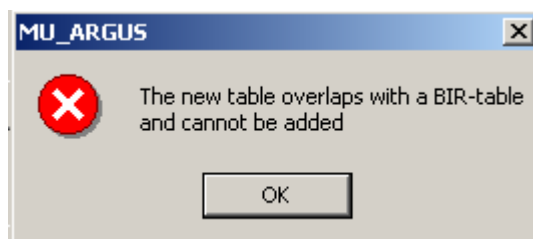


Figure 6. Selection of two sets of key variables

If overlapping risk tables are defined, i.e. if the user specifies two or more sets of key variables with at least one common variable, μ -Argus returns an alert message such as the one shown here. After that, either the introduction of an additional risk table is inhibited, or μ -Argus forces the user to remove the overlapping table(s). Recall also that when multiple risk tables are defined, the button **Set table for risk model** has to be used for every single table.



To start the procedure that computes the risk table and close the **Select Combinations** window, press the button **Calculate tables** (indicated by number **4** in Figure 5) . The software runs the routine that computes the risks and shows an **output window**.

7.2.1. Output window

Being common to any procedure run in μ -Argus, this window (see for instance Figure 7) is not directly connected with the risk methodology. Indeed it reports the number of unsafe combinations computed according to the *traditional (Dutch) methodology* which relies on the concept of uniques or fingerprints. Recall that in the traditional framework the number of unsafe records depends on the *frequency threshold*: a combination is unsafe if its frequency is below a given threshold value, that is imputed in the **threshold** box of the **Select Combination** window (see Figure 5, number 5). Note that the settings of the **Select Combinations** window would apparently allow to adopt *both* the traditional (Dutch) approach based on the frequency threshold *and* the individual risk approach. Indeed even when the individual risk methodology is adopted, the frequency **threshold** box (see Figure 5, number 5) is active. However in this case the software does not make use of the selected threshold value, as it has no role in risk computation: any value of the frequency threshold has the same effect on risk computation, although the standard output given by μ -Argus (shown in Figure 7) changes according to the frequency threshold.

The output also differs according to whether one or more risk tables have been specified: in either case the left panel shows the **number of unsafe records** (defined according to the traditional

methodology) for each **dimension**, e.g. level of cross-tabulation of the key variables. As already explained, the default output does not return information specific to the risk model, nonetheless it can be useful for a generic assessment of the file. For instance, screening the left panel can suggest which variables may undergo recoding, whereas the right panel may help identifying the categories to be collapsed. In our example, when all four key variables are selected into one single risk table, the output contains the number of unsafe records for each dimension up to 4 (Figure 7).

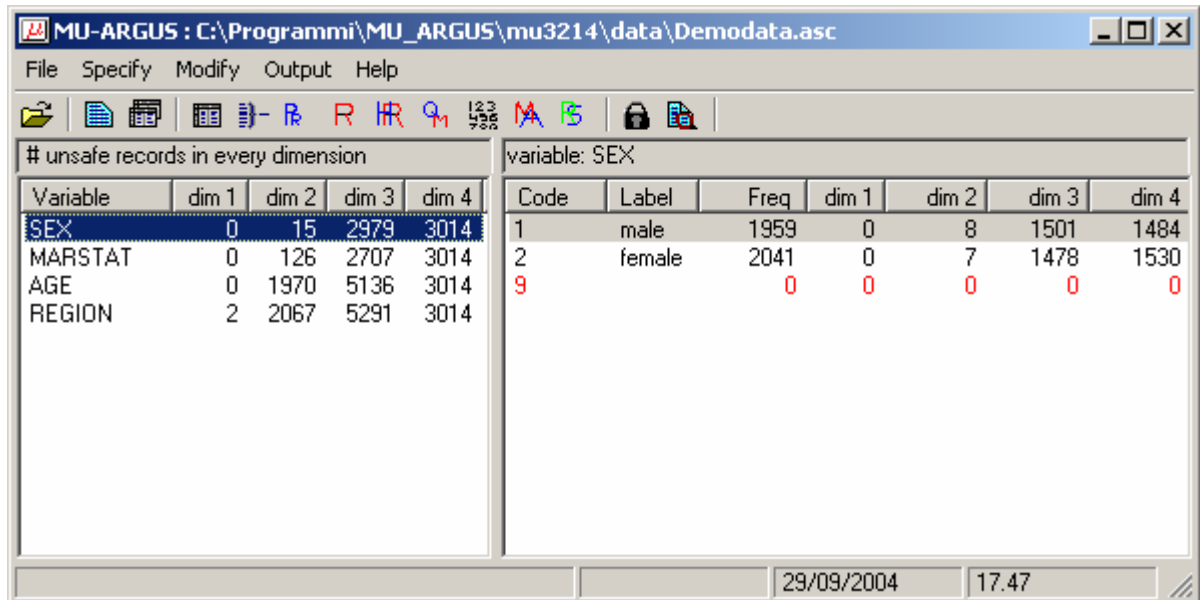


Figure 7. Output window for the individual risk methodology (4 key variables)

When we define two separate risk tables, each having two key variables, a single output window (see Figure 8) shows both tables. Compare the output shown in Figure 7: although the key variables are the same in both windows, in Figure 8 combinations up to dimension 2 are shown. In the latter case the two tables are indeed considered separately. Clearly the highest dimension considered in the output window depends on the maximum level of cross-classification allowed in each of the risk tables defined by the user.

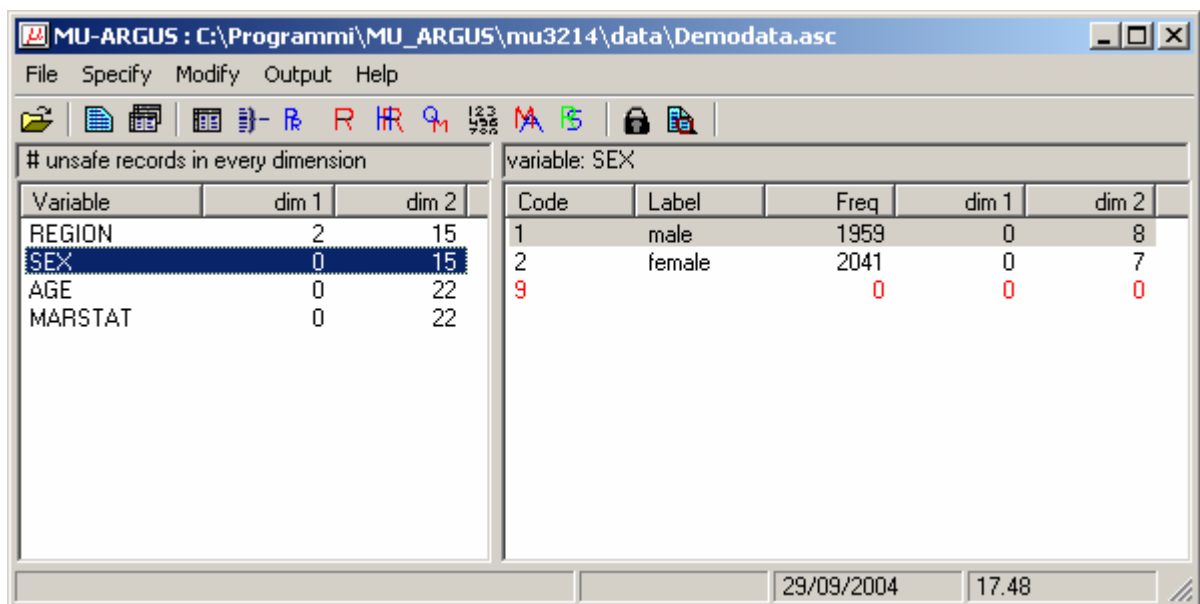


Figure 8. Output window for the individual risk methodology (2 sets of 2 key variables)

7.3. Menu Modify|Risk Specification: the Risk Chart window

As already mentioned, the risk table stored in μ -Argus contains the individual risk for each combination of key variables. This table cannot be accessed, but several aspects of the risk structure of the data can be inspected on request, either by selecting the menu **Modify|Risk Specification**, or by clicking the **R** icon on the toolbar. The **Risk Chart window** contains the available information on the risk structure of the file. If global recoding is applied to any of the key variables, the risk table is updated, and for assessing such protection procedure users must rely on a new inspection of the Risk Chart window.

7.3.1. Risk histogram

First of all, the **Risk Chart** window (Figure 9) illustrates the histogram of the individual risks, presented on a logarithmic scale. Note that if more than a single risk table is specified, users are requested to choose the graph to be inspected first.

The key variables used to compute the risk are specified on top of the graph. Finally, the option **Cumulative chart** transforms the histogram into a cumulative frequency graph.

7.3.2. Risk levels in the data before suppression

The risk chart window also contains information that refers to the file *before suppression* is applied to the data. The left-hand panel (**Maximum levels in the file**) shows the **maximum individual risk** in the file (5) and its **re-identification rate** (6), expressed as $100 \cdot \xi \%$. The first of these quantities gives the highest individual risk that is attained in the original file, whereas the second represents the global risk of the file, i.e. the percentage (over the file size) of expected re-identifications in the file, would the file be released *without suppressions*. If global recoding is applied, this information is updated and the above mentioned values indicate the extent to which recoding affects the data. These values can therefore be used to assess whether the recoded file can be released without suppressions.

7.3.3. Setting the risk threshold to be used for local suppression

Before describing the rest of the window, we remark that, under the individual risk methodology, local suppression is driven by the value of the individual risk. This means that a rule, which is based on the risk, is applied to identify the *unsafe* records, i.e. the ones that need to be protected by suppression. Here the **unsafe records** are defined as those records whose risk is above a given **threshold**. Specification of this latter parameter is left to the user, who can proceed either directly, by selecting a **threshold** value for the **individual risk** r , or indirectly by imposing either a threshold (i.e. a maximum tolerable level) on the **re-identification rate** $100 \cdot \xi$, or a target number of **unsafe records** n_u . In any circumstances, the left-hand panel, that contains information relative to the risks of the unprotected file, should be used as a benchmark in threshold setting.

First of all, the user can select the individual risk threshold manually, by using the **slider** (1) below the graph (see Figure 9). The **risk threshold** box 2, the corresponding **number of unsafe records** 3 (i.e. the records whose risk is above the selected threshold) and finally the **threshold** on the **re-identification rate** (4) are automatically updated. A risk threshold can also be selected by using the tools in the **threshold setting** area of the risk chart window (central panel). If the user has already in mind a threshold on the individual risk, this value can be typed in directly in the **Risk threshold** box 2. Clearly this value should be smaller than the maximum risk in the file, which is shown in 5. As discussed in Section 4.1, the individual risk threshold can also be set through the re-identification rate; to this aim, the user must type a value in the **re-identification rate threshold** box 4. Clearly this value cannot exceed the re-identification rate shown in the left hand panel (6). In all cases, the **Calc** button relative to the current box updates the quantities in the other boxes

(number of unsafe records and re-identification rate threshold/risk threshold, respectively).

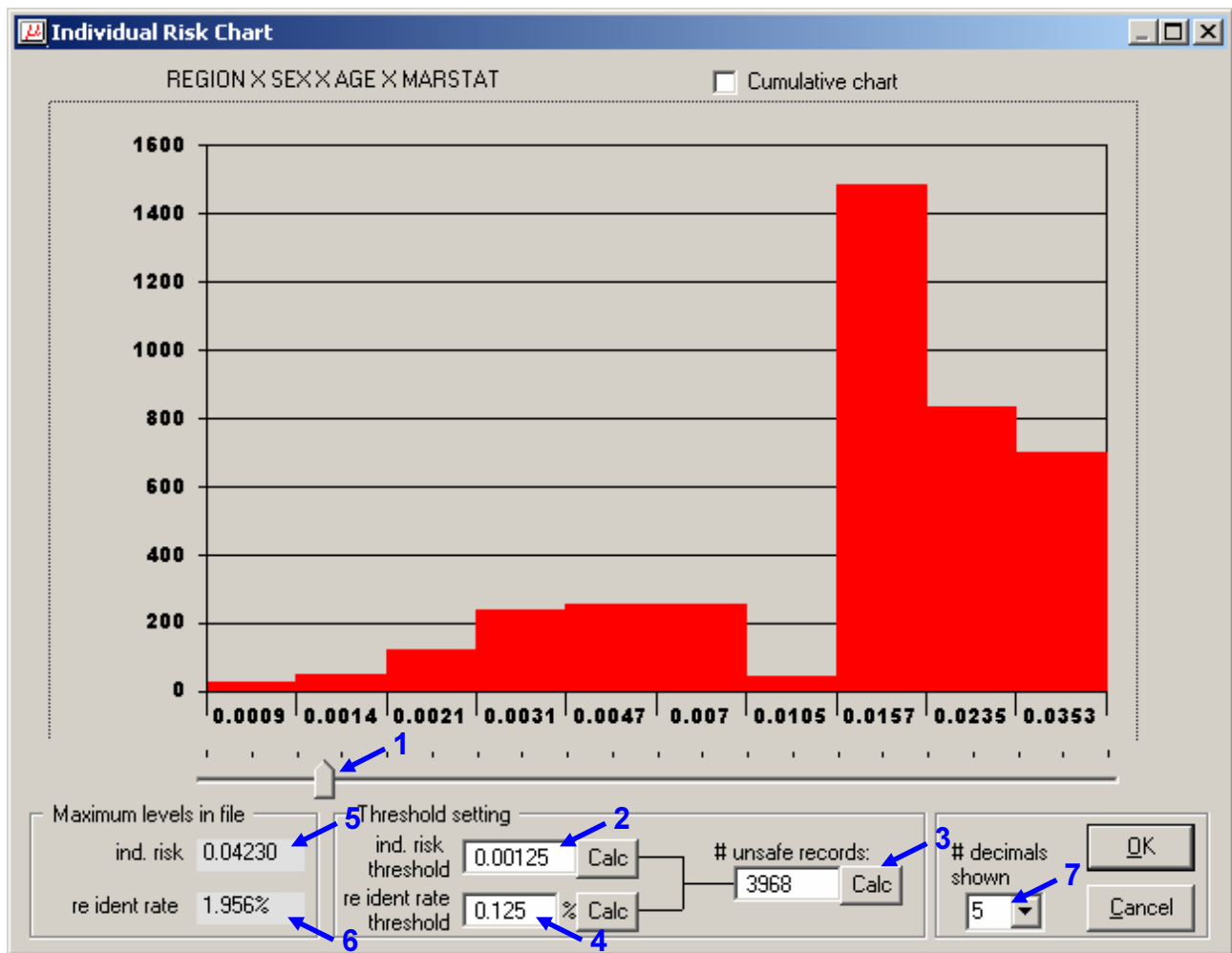


Figure 9. Risk chart for the set of key variables Region, Sex, Age and Marital status

For a given risk threshold r^* specified by the user, μ -Argus computes the corresponding number of unsafe records as $n_u = \text{Card}\{i: r_i \geq r^*\}$. However, as **Fout! Verwijzingsbron niet gevonden.** shows, the estimated risks in fact form a discrete set of values; for this reason there is a whole set of threshold values compatible with the same number of unsafe records. Once typed a threshold r^* and determined the corresponding n_u , the user should compute the *actual* threshold level, e.g. the maximum *estimated* risk compatible with the given number n_u of unsafe records. This is obtained by pressing the **Calc** button **3**.

Of course the user can also determine the thresholds **2** and **4** on the individual risk and the re-identification rate, respectively, that correspond to the desired number of unsafe records; this must be typed in **3**. If necessary, the precision shown in the threshold boxes can be changed by selecting the appropriate number of digits in the right hand-side of the window (**7**).

Pressing the button OK sets the individual risk threshold at the *current* level. Recall that when the suppressed file is built, this level will be used to identify the unsafe records, i.e. the records to be protected. Choice of the threshold depends on the assumed scenario and on the uses of the file, e.g. whether it is intended as a microdata file for research or as a public use file. A subjective judgement of the availability of the external archive can also affect the threshold level. Eventually, the assessment of the appropriate threshold level for the data at hand, relative to the assumed scenario, is left to the person who is responsible for the release. In our DEMODATA example, it could be considered safe to release a data file in which the expected number of re-identifications is certainly less than 5 expected re-identifications out of 4000 records. This corresponds to a threshold

100 ζ^* % = 0.125 %. μ -Argus calculates a corresponding risk threshold $r^* = 0.00125$, which means that in the released data the maximum probability of re-identification will be less than 1 out of 800 records¹. However, prior to protection, the records whose risk is above this threshold are 3968; these are marked by the software as unsafe and undergo local suppression: applying suppression with this threshold would lead to a file in which nearly all records have at least one missing value, which cannot be considered a satisfactory compromise between protection and information loss.

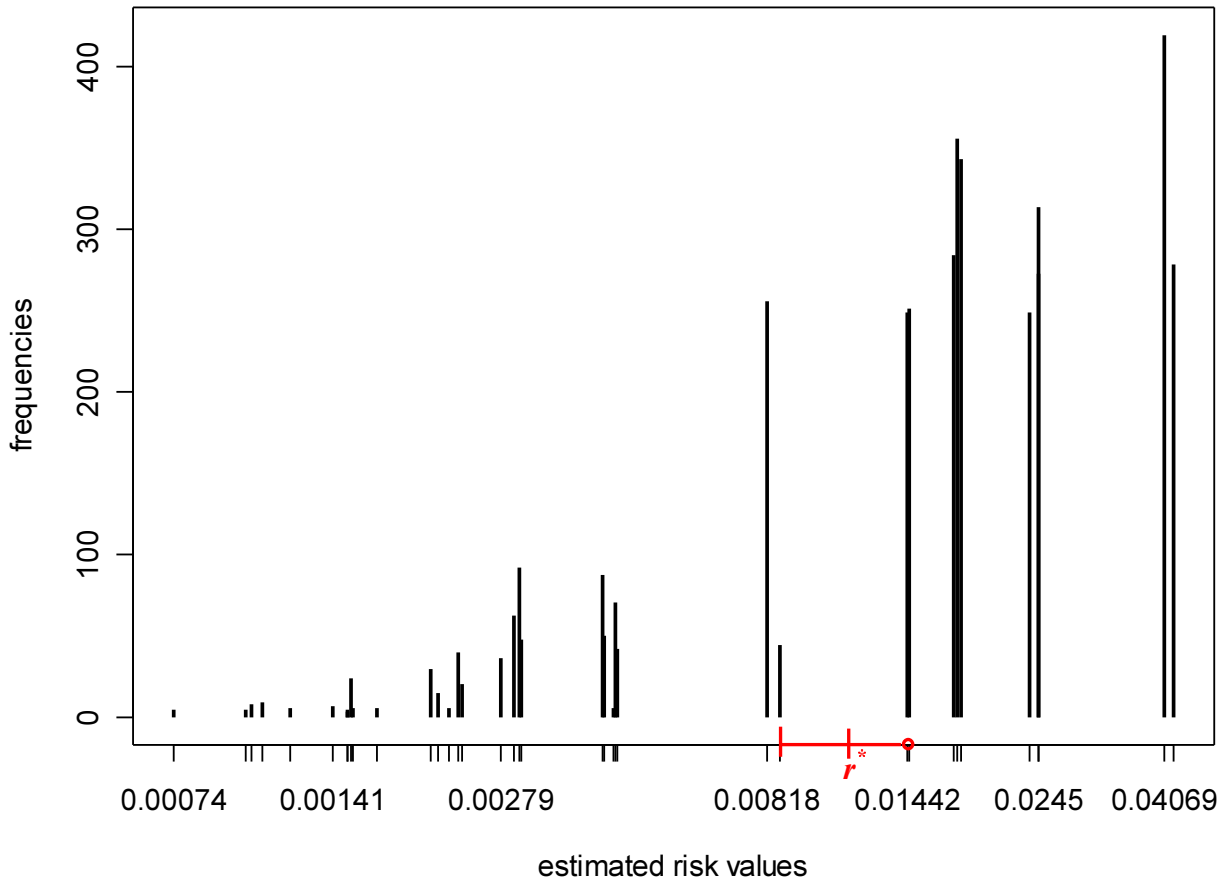


Figure 10. Graph of the risks (on a logarithmic scale) estimated for the DEMODATA file (key variables: Region, Sex, Age and Marital status; the interval of threshold values returning the same number of unsafe records as r^* .

As the above example shows, care should be taken of the number of unsafe records induced by the selected risk threshold. An exceedingly high number of unsafe records implies a high number of suppressions, that is, a highly corrupted microdata file. To ensure quality of the released data, it is advisable to apply *global recoding* and/or *bottom/top-coding* first. For the same example as in Figure 9 above, we recoded Age and Region to 6 and 18 classes, respectively. To this aim, the recode files REGION.GRC and AGE.GRC accompanying the μ -Argus DEMODATA file were used. Recall that after any manipulations of the key variables, the Risk Chart window is automatically updated; the Risk chart shown in Figure 11 can be accessed via the **R** icon, for instance.

¹ In our example the *actual* risk values (i.e. the closest observed individual and global risk values that are below the nominal level) are 0.119 % for the re-identification rate threshold and 0.00119 for the individual risk threshold; higher protection is therefore achieved. To determine the actual levels, set the desired threshold, update the other values and use the calc button (3) of the number of unsafe records. In order to avoid that the number of unsafe records changes unexpectedly, merely as the effect of rounding, do not press the **Calc** button 2 again at this stage. Indeed in that case the software re-reads the value appearing in the risk box *using the digits shown*; this implies some rounding that might affect the number of unsafe records, that might change even if the risk threshold was not changed by the user.

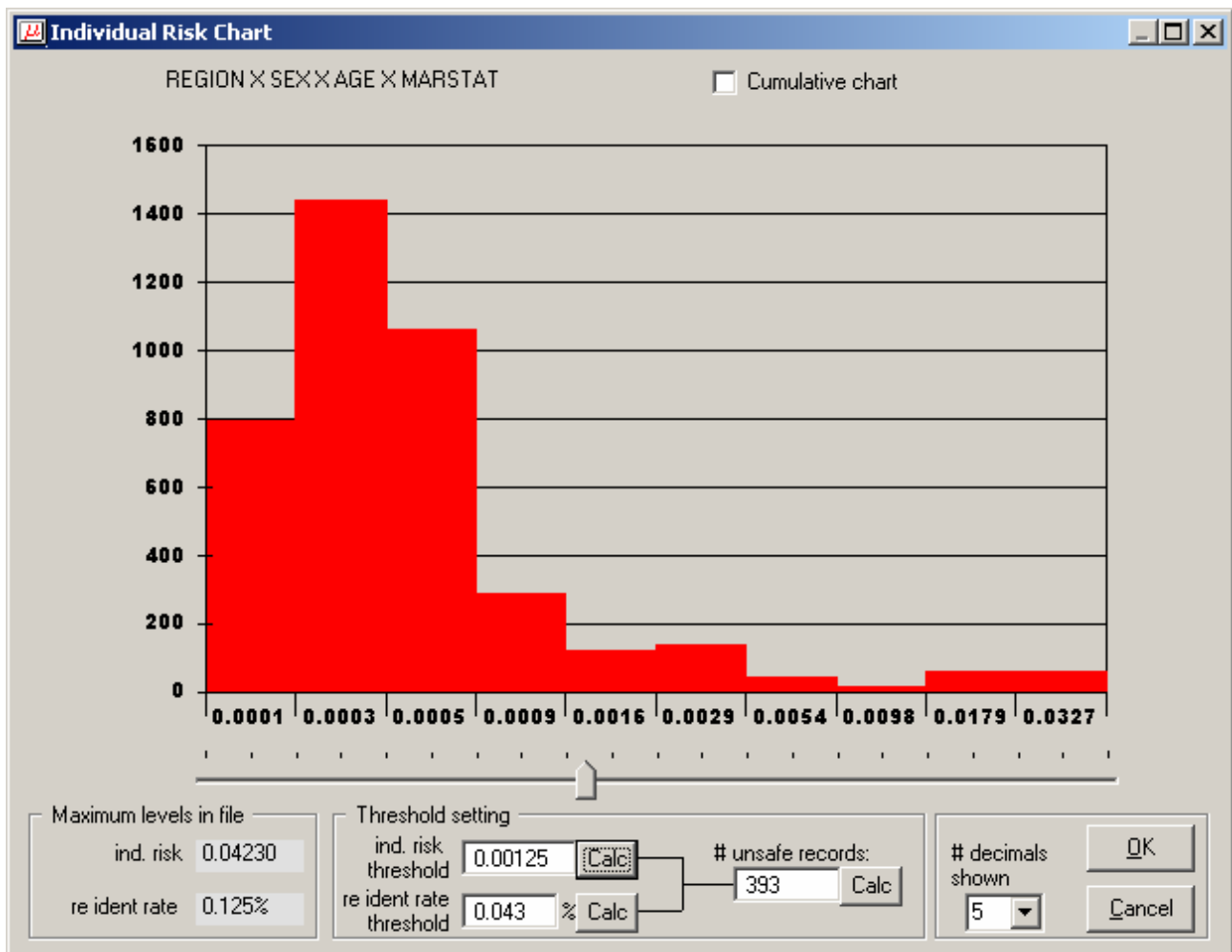



Figure 11. Risk chart for the set of key variables Region, Sex, Age and Marital status after recoding is applied to variables Region and Age.

Comparison of Figure 9 and Figure 11 illustrates the changes induced in the risk distribution by global recoding. Setting the same risk threshold as before, we notice that recoding has considerably lowered the number of unsafe records, that are now less than 400. Of course, the decrease in the number of unsafe records depends on how recoding is applied, i.e. on the number of collapsed categories and their frequencies. Compared to the previous example, records are now allocated to cells of a broader contingency table; consequently the sample cell frequencies are increased and the risk is lowered for most records, not only the unsafe ones. The global risk of the unprotected file has also sensibly decreased (0.125%, or 5 expected re-identifications in the file, instead of 1.956%, i.e. about 78 expected re-identifications in the raw data).

In practice a combination of approaches is usually adopted, as the example has made clear: so far, Istat has been applying first global recoding; the re-identification rate of the file before suppression, the maximum risk level and the whole individual risk graph help the user in assessing whether the solely recoded file can be considered safe; current practice at Istat consists in first recoding the variables to the finest detail at which the estimates retain their planned precision. Some of the key variables –usually geographical variables and age- are also considered for further recoding. Numerical key variables may also be treated by top/bottom coding. If the risk properties of the file do not meet the required standard, local suppression, based on the individual risk of re-identification, is applied. Clearly the number of suppression induced by the threshold should always be inspected to decide whether other solutions (further recodings, suppression of variables from the file...) should be considered. Such a strategy aims at maintaining the quality of the file, while guaranteeing confidentiality.

7.4. Menu Modify|Household Risk Specification: the Household Risk Chart window

When a household identifier is detected in the metadata, and the risk methodology is chosen, the household risk described in Section 5 is automatically computed. The **Modify** menu (or the  button) permits to access the household risk window. Its aspect is very similar to that of the Individual Risk Chart window.

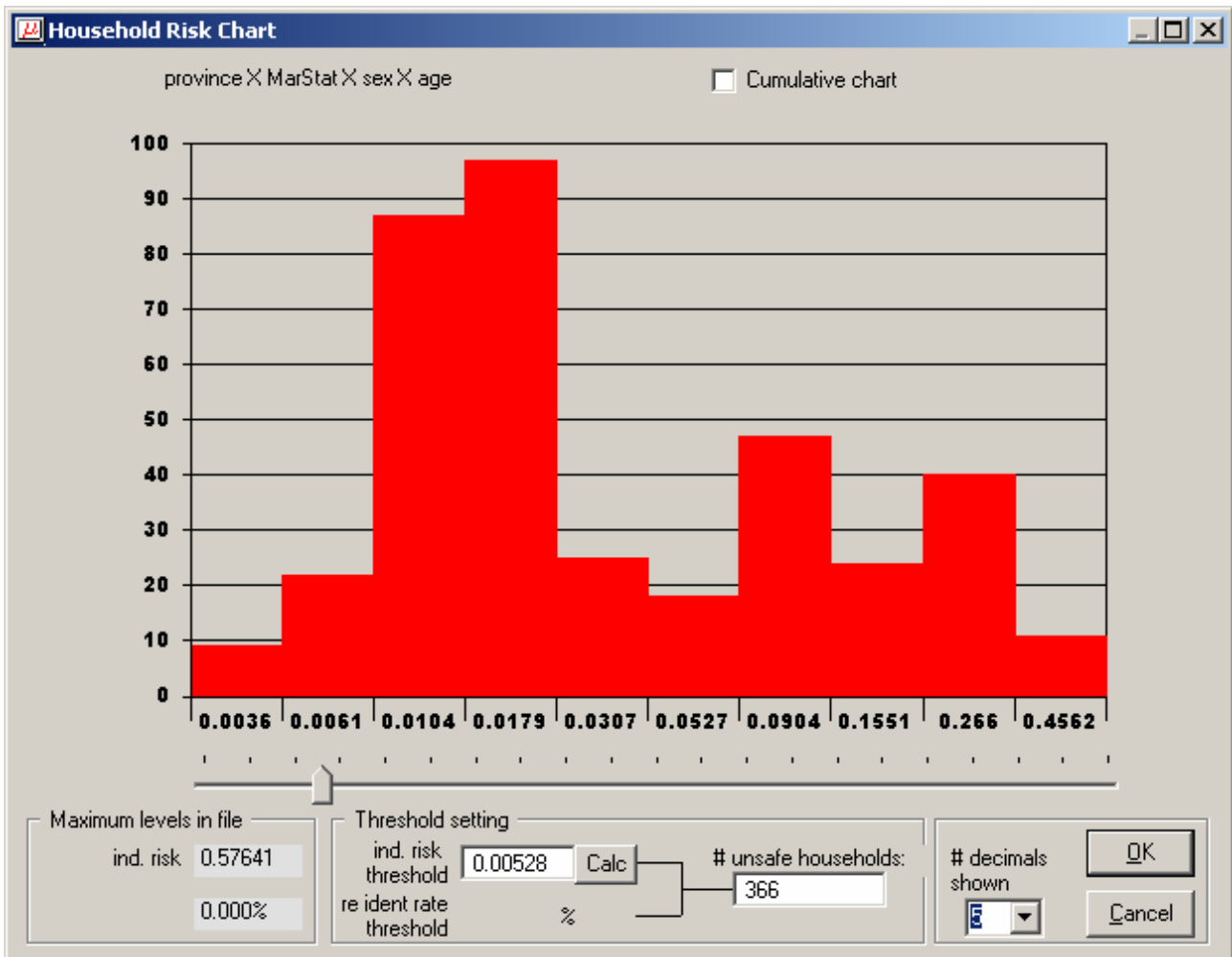


Figure 12. Household risk window for the HHDATA demonstration file

Using the demonstration file SAMHH.ASC, we follow the usual procedure to create a risk table having the following key variables: province, MarStat, sex, age. The corresponding household risk chart window is shown in Figure 12. By this window the user can perform an assessment of the (household) risk of the data and identify the unsafe records as discussed in the next section.

7.4.1. Setting the household risk threshold and identifying unsafe records

In a file having a household structure, the unsafe *records* are defined in two steps: first a threshold on the *household* risk is selected; in this way *households* are defined as safe or unsafe. However, depending on the values of the key variables, within the same household there will be high risk as well as low risk records. As the household risk is the composition of the individual risks of the members of the household, to decrease the household risk it will be most effective to protect the records whose individual risk is highest. Therefore the second step in the procedure is to define a rule that permits to identify, within the unsafe households, the records to be suppressed. These will be referred to as the unsafe (or at risk) records. By this rule, in fact only a *subset* of the records

belonging to unsafe households will be classified as unsafe. As discussed in Section 5.2, our rule to detect the unsafe records relies on a condition that is sufficient to ensure that, after protection, a household is safe for a given threshold r^{h*} on the household risk. Such condition is that all records in the household do not contribute more than $(1/d)$ -th of the household risk, d being the household size. This in fact amounts to using different thresholds for different household sizes.

To illustrate the procedure to identify unsafe records, we refer to the SAMHH data example. *For illustration purposes only*, we position the slider at a threshold $r^{h*}=0.13431$. This value identifies 71 unsafe households, three of whom are shown in Figure 13. Here for instance the first household has three members, of whom only the second has high individual risk. For a household risk threshold $r^{h*}=0.134310$ as in the example, records of the first household are defined *safe* whenever their individual risk is below 0.04477; only the second record in the first household is therefore set to unsafe. Likewise, in the third household, the first two records are at risk for the given threshold $r^{h*}=0.134310$. For households of size 1, like the third household in Figure 13, this procedure induces no difference between the household and the individual risk threshold.

742221	32421	162	42	0.040238671962	0.199476456345
742221	11192	0	42	0.146135380971	
742221	12162	0	42	0.023163578416	
751111	31882	0	88	0.238164748136	0.238164748136
761111	21331	572	33	0.075354624319	0.198838280708
761111	22331	110	33	0.075354624319	
761111	12	72	0	0.031978745825	
761111	11	52	0	0.031978745825	

Figure 13. Data selected from the demonstration file samhh.dat. Each records is sided by the individual and household risk. The household risk is reported only for the first record in the household.

Once a household risk threshold r^{h*} is selected in the Household Risk Chart window, the number of unsafe households and records, respectively, is shown in the proper box. As explained in the previous example, the number of unsafe records is generally smaller than the total number of records belonging to unsafe households.

7.4.2. Validity of the assumptions that define the scenario for estimating the household risk

The disclosure scenario defined in Section 5.1 for the release of files of households assumes that the intruder's archive refers to individuals and does not contain information about the household structure. Depending on the information provided by the public archives, sometimes it might be more realistic to assume that the re-identification attempt also makes use of household variables such as the size of the household and/or the household type. In this circumstance we recommend to use these as additional key variables in computing the risk, so as to allow for the additional information concerning the household. If not directly available in the data, both variables can be computed from the file.

If instead it is deemed more appropriate to define a scenario under which the intruder has access to an archive of households, each containing as many individual records as there are members in the household, then the definition of household risk is not fully appropriate. In this circumstance, including the household size and the household type in the key variables permits to approximate the appropriate household risk, i.e. allows for part of the dependence between records in the same household.

7.5. Output: Make protected file window

Once the individual and/or global risk thresholds, and the corresponding number of unsafe records are deemed appropriate, the output file (**safe file**) can be produced. From the menu **Output** select

Make protected file. The window **Make protected file** pops up (see Figure 14). This window looks the same for all the protection methods offered by μ -Argus. In the application of the individual risk methodology, the left panel shows the key variables only; as partial suppression only affects these variables. If this is reasonable, users can ask for no suppressions, otherwise a criterion to place missing codes in the data must be chosen. μ -Argus has two options, either **entropy** or **suppression weights**. With the latter option, users can adjust the suppression weights manually as it happens with the traditional (Dutch) technique. Introducing different weights can be useful to give different suppression penalties for different key variables: for instance, users can assign sex the maximum weight (100) so as to discourage suppressing that variable, as long as it is possible. With both options, the suppression criterion used by μ -Argus is based on an optimisation algorithm. Missing codes are placed so as to minimise a measure of information loss (either entropy or suppression weight). In any case, the effect of suppression is to lower the individual risk of the protected records; after protection, this will be below the selected threshold r^* .

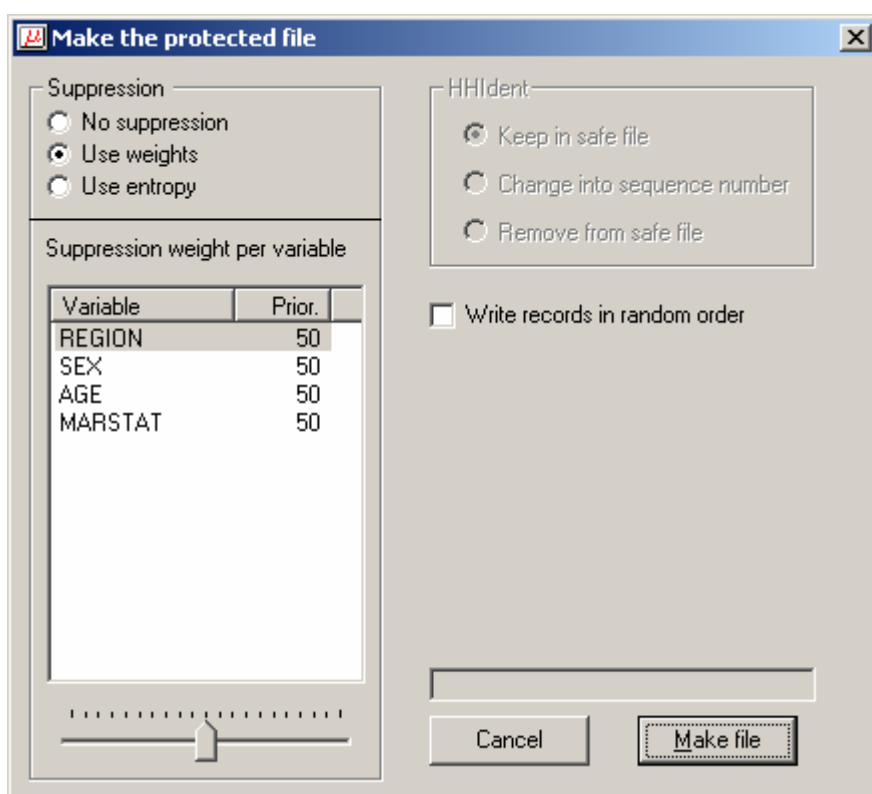


Figure 14. Make protected file window

Pressing **Make file** will store the safe file. A report is produced, that details all the procedures that lead to the safe file. Besides information on the original and protected file, the report details the main aspects of the protection: a list of the key variables used in the model and the risk table, the list of recodings, when global recoding has been applied, detailed for each recoded variable, the maximum risk levels in the output file, i.e. the individual risk threshold, the global risk threshold, and, if applicable, the household risk threshold; finally, the marginal number of suppressions per variable needed to produce a data file that is safe for the given threshold levels.

The report also contains a record description of the output. We next show an excerpt from one such report for the example at hand (for the DEMODATA file, that does not carry a household structure).

μ-ARGUS Report

Safe file created date: 10-10-2004 time 13:06:58

Original file:	C:\Programmi\MU_ARGUS\data\Demodata.asc
Original meta file:	C:\Programmi\MU_ARGUS\data\Demodata.rda
Number of records:	4000
Safe data file:	C:\Programmi\MU_ARGUS\data\demodata.saf
Safe meta file:	C:\Programmi\MU_ARGUS\data\demodata.rds

Identifying variables used

Variable	No of categories (missings)	Household var
<i>REGION</i>	18 (2)	
<i>SEX</i>	2 (1)	
<i>AGE</i>	6 (2)	
<i>MARSTAT</i>	4 (1)	

Frequency tables used

Treshold	1	2	3	4
1	<i>REGION</i>	<i>SEX</i>	<i>AGE</i>	<i>MARSTAT</i>

Global recodings that have been applied:

REGION

Code	Categories
1	1-10
2	11-20
3	21-30
4	31-40
...	
16	151-160
17	161-170
18	171-
98,	Missing 1
99	Missing 2

AGE

Code	Categories
1	-20
2	21-30
3	31-40
4	41-50
5	51-60
6	61-
8	Missing 1
9	Missing 2

Base Individual Risk has been applied:
table: REGION x SEX x AGE x MARSTAT
Ind. risk: 0.001250
Ind. re-ident rate: 0.000431
HH-risk: 0.000000

No other modifications

Suppression overview

Name	Suppr. Weight	Number of suppressions
REGION	50	375
SEX	50	0
AGE	50	11
MARSTAT	50	7
KINDPERS	50	0
NUMYOUNG	50	0
...		
ADDJOB	50	0
JOBFIND	50	0
Total	-	393

Record description safe file

Name	Starting pos	Length	Decimals
REGION	1	3	0
SEX	4	1	0
AGE	5	1	0
MARSTAT	6	1	0
KINDPERS	7	1	0
NUMYOUNG	8	1	0
...			
ADDJOB	40	1	0
JOBFIND	41	1	0
WEIGHT	42	6	2
INCOME	48	8	0
ASSETS	65	5	0
DEBTS	72	4	0

The final results can be assessed by inspection of the **suppression overview** in the report above. If the number of suppressions per variable is not satisfactory, the user may want to *recode* selected variables; in this case, go to **Modify|Global Recode** and inspect the new Risk Chart, possibly selecting a new risk threshold. The process may be iterated until satisfactory results are produced. The risk threshold, *if appropriate*, can also be modified; in this case, go back to the Risk Chart.

References

Abramowitz, M. and Stegun, I. A. (1965). Handbook of Mathematical Functions, Dover, New York.
 Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination, *Pre-proceedings of New Techniques and Technologies for Statistics*, 1, 225-232.

- Benedetti, R. and Franconi, L. and Capobianchi, A. (2003). Individual Risk of Disclosure Using Sampling Design Information. *Contributi Istat* n.14-03.
- Bethlehem, J., Keller, W. and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association* 85, 38-45.
- Carlson, M. (2002). Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition* 5, 901–925.
- Chen, G. and Keller-McNulty, S. (1998) Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* 14, 79–95.
- Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association* 87, 367–382.
- Di Consiglio, L., Franconi, L. and Seri, G. (2003). Assessing individual risk of disclosure: an experiment, *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg.
- Duncan, G.T. and Lambert, D. (1986). Disclosure-limited data dissemination (with comments), *Journal of the American Statistical Association* 81: 10-27.
- Elamir, E. and Skinner, C. (2004). Modeling the re-identification risk per record in microdata. *Technical report*, Southampton Statistical Sciences Research Institute, University of Southampton, U.K.
- Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14, 385-397.
- Franconi, L. and Poletini, S. (2004). Individual risk estimation in μ -argus: a review. In: Domingo-Ferrer, J. (Ed.), *Privacy in Statistical Databases*. Lecture Notes in Computer Science. Springer, 262-272
- Hundepool, A. and van de Wetering, A. (2004). μ -Argus user's manual v. 4.0, CASC Deliverable 2-D2
- Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* 9, 313–331.
- Poletini, S. (2003a). A note on the individual risk of disclosure. *Contributi Istat* n.14/2003
- Poletini, S. (2003b). Some remarks on the individual risk methodology. In: *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg.
- Poletini, S. and Stander, J., 2004. A bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In: Domingo-Ferrer, J. (Ed.), *Privacy in Statistical Databases*. Lecture Notes in Computer Science. Springer, 247-261
- Rinott, Y. (2003). On models for statistical disclosure risk estimation. In: *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg.
- Skinner, C.J. and Holmes, D.J., (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 4, 361-372.
- Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, 111, New-York: Springer Verlag.
- Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, New York: Springer-Verlag.