Working Paper No. 17
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (v): Risk assessment

ASSESSING INDIVIDUAL RISK OF DISCLOSURE: AN EXPERIMENT

Invited paper

Submitted by the National Institute of Statistics (ISTAT), Italy [1]

[1] Prepared by Loredana Di Consiglio, Luisa Franconi (franconi@istat.it) and Giovanni Seri).

Assessing individual risk of disclosure: an experiment

Loredana Di Consiglio, Luisa Franconi and Giovanni Seri
Istat, Servizio della metodologia di base per la produzione statistica,
Via A. Depretis 74/B, 00184 Roma, Italy

1. Introduction

When releasing microdata file, assessing the possibility for external users of performing confidentiality breaches should be considered a good practice. The assessment can be carried out by defining a model for disclosure risk and quantifying such risk for the release of the microdata. This will allow the person who is in charge of the release to make a responsible choice on the basis of statistical evidence and, if the observed risk is considered not tolerable, to establish effective protection measures to be put in place.

In social surveys, the risk is usually defined as a function of combinations of scores on variables present in the microdata file to be released that are also publicly available (such as sex, age etc.). The key issue in risk assessment is how to discriminate between situations at risk (for examples, sample uniques that represent combinations that are very rare in the population) and cases that do not present problems (e.g. sample uniques that represent common features in the population). Although the theme of risk assessment is quite recent a vast literature is present; recent examples include Skinner and Elliot (2002), Fienberg and Makov (2001).

In this paper we conduct an experiment to assess the performance of the individual risk of disclosure initially proposed by Benedetti *et al.* (1999) which is currently in the process of being implemented in the software μ-Argus as part of the CASC project. The aim is that of investigating 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. To do this we concentrate on a subset of the 1991 Italian population census and select six hundreds samples using the Labour Force Survey design in order to replicate, as much as possible, the real situation.

In Section 2 we briefly introduce the individual risk model. Section 3 describes the structure of the experiment and the technical detail of the sampling design used for the Labour Force Survey. Section 4 presents some of the results of this experiment whereas Section 5 contains the conclusions and suggestions for further work.

2. The individual risk model

We give a very brief overview of the individual risk model used in this experiment; for further information and details see Benedetti *et al.* (1999), Benedetti and Franconi (1998) and recent development in Polettini, (2003). The aim of the individual risk model is to estimate a risk of disclosure for each unit i in the random sample to be released. The assessment of the risk then allows to select the units that have to undergo a protection: these will be all the units that present a risk higher than a predefined threshold that represents the maximum tolerable risk (see Polettini, 2003 on ways to choose it).

The definition of disclosure in use is the re-identification disclosure (Duncam and Lambert, 1986). The way in which an intruder may identify a unit is by mean of the key variables, variables that allows identification and are publicly available. In social surveys these key variables are categorical and this fact allows to switch the attention from the units to the corresponding combinations of categories of the key variables $k=1, \ldots, K$. Let f_k and F_k be, respectively, the number of units in the random sample and the number of individuals in the population with the k-th combination of categories of the key variables; F_k is

unknown for each k. In the sample to be released only a subset of the total number K of combinations is observed and only this subset, for whom $f_k>0$, is of interest to the disclosure risk estimation problem. The intruder tries to link unit i in the sample to an individual in his register by comparing the values of the key variables. An identification occurs when based on this comparison, an individual i^* in the register is chosen as a match to i and this link is correct. The individual risk for unit i in the sample, as far as this experiment is concerned, is defined as r_i=P(i correctly linked to i^* | sample) and is given by, see Benedetti *et al.* (1999):

$$\sum_{h \geq f_k} \frac{1}{h} P(F_k = h \mid f_k) \stackrel{\text{def}}{=} \hat{r}_i \qquad (1)$$

In order to estimate the individual risk of re-identification, Benedetti and Franconi (1998) assume that $F_k|f_k$ is distributed according to a negative binomial distribution with success probability $p_k = f_k/F_k$ and number of successes f_k. The risk (1) can be seen as $E(\hat{F}_k^{-1}|f_k)$ and using the analytic expression for this expectation under the negative binomial distribution it is possible to evaluate the risk (see Benedetti *et al.* (1999) and also Polettini, 2003).

A crucial step in the procedure is the estimation of the parameter p_k. Benedetti and Franconi (1998) propose to estimate it by:

$$\hat{p}_k = \frac{f_k}{\sum_{i \in k(i)} w_i} \qquad (2)$$

where w_i is the final weight attached to each unit in the sample and the sum is over all units i who share the same combination $k=k(i)$. The experiment described in Section 3 is set up to investigate the effectiveness of the estimates of the individual risk model.

3. The experiment

The idea is to simulate a situation that would resemble as much as possible a real survey from a real population. To do this we have considered the 1991 Italian Population Census data from 4 administrative Italian regions (Val D'Aosta, Veneto, Lazio and Campania) as the source of information for the experiment. The total number of individuals in the population from these four regions is over 15 millions individuals (15,142,320). Then we consider the largest and most important survey in Italy: the Labour Force Survey (LFS). This survey is repeated four times in a year; the overall sample size equals about 72,000 households (approximately 200,000 individuals) in each quarter. Accordingly, each sample from the four selected regions amounts to roughly 18,000 households (approximately 54,000 individuals). For this experiment six hundred samples have been selected according to the LFS sampling design which is described in details in the following section.

The data set available from the 1991 census contained the following variables: sex (2 categories), age in years (from 0 to 110), region of residence (4 selected regions in this study), position in profession (14 categories) and relationship with the head of the household (13 categories). These are the variables we inspected as key variables. Note that estimates from the LFS are significant at regional level; for this reason the geographical information of the Microdata File for Research routinely released from this survey is the administrative region and we analyse only this geographical area.

The disclosure scenario which is at the base of the individual risk model makes the hypothesis that an intruder has the whole of the population and tries to link a unit in the sample to one individual in the population. Accordingly, for the intruder, the probability of linking one unit in the sample to one in the population is equal to $1/F_k$ where k is the corresponding combination to which the unit belongs to. We call this quantity the *real risk, R.* One of the aim of this simulation study is to analyse the effectiveness of the protection procedure based on the estimates of the individual risk. Some of the results of this experiment are presented in Section 4.

3.1 *The LFS sample design*

The LFS sampling design is a two-stage sampling design with stratification of the municipalities, the primary sampling units (PSU), and systematic selection of the households, the Secondary Sampling Units. In each province (administrative areas inside the region), the PSUs are grouped into two area types according to their dimension in terms of population: the Self-Representing Area (SRA), consisting of the larger municipalities, and the Non Self-Representing Area (NSRA), consisting of the smaller ones. The classification is based on a function of the minimum number of interviews to be carried out in each municipality and the sampling ratio, f. The value of f is a constant for all the provinces in a region but varies across regions.

Each municipality in the SRA represents a single stratum and is included in the sample with probability one. In the NSRA, after having sorted the municipalities by their dimension, they are divided into strata of approximately constant total number of residents. Two PSUs from each strata are selected without replacement and with probability proportional to their size using the randomized systematic procedure that was first introduced by Madow (1949).

In each sampled municipality, the households are sampled by means of systematic sampling. All members of each household in the sample are interviewed.

The final weight of each household is derived from the basic weight (see below) by means of the calibration process (Deville and Särndal, 1992) in order to preserve the known population totals. Such auxiliary information is usually provided by means of demographic surveys; in this simulation they are obviously known. In the LSF the population totals which are maintained are the population of each region by sex and fourteen age-classes and the population of each province by sex. For simplicity in our simulation we have reproduced the constraints only at regional level. In particular, let N_h be the number of municipalities, P_h the number of persons, n_h the number of sample municipalities in stratum h; M_{hb} the number of households, P_{hb} the number of persons, m_{hb} the number of sample households in municipality b of stratum h. The basic weight of household j of municipality b in stratum h, K_{hbj} equals the inverse of the probability of inclusion, i.e. (see Cochran, 1977) $K_{hbj} = \frac{P_h}{n_h P_{hb}} \frac{M_{hb}}{m_{hb}}$. Note that for the municipalities in the SRA n_h=1 and P_{hb}=P_h, so $K_{hbj} = \frac{M_{hb}}{m_{hb}}$. Using sex and fourteen age-classes (0-14, 15-19,…,70-74, 75-) as auxiliary information, the final weight, K'_{hbja}, for unit l in sex-age class a of household j of municipality hb is given by $K'_{hbja} = K_{hbj} \frac{P_a}{\sum_{hbjl} K_{hbj} d_{hbjl}}$, where d_{hbjl} is an indicator function equal to 1 if unit $hbjl$ is in class a and 0 otherwise and P_a is the total number of units of one region in sex-age class a.

The sampling scheme used for the LFS is common to many other social surveys and represents a standard sampling design at Istat.

4. Results

In this section we present some of the results of the study. In Section 4.1 we examine various comparisons of the true and estimated individual risk in some particular samples; this is to investigate the behaviour of the individual risk under different conditions. The estimate of the individual risk for unit i is a function of the frequency of the corresponding combination k in the sample, f_k and the final weights, w_i for units i in combination k. The different conditions then refer to different values of the frequencies f_k, different sets of weights and different key variables. In particular, we analyse the individual risk behaviour in the presence of different relationships between the set of variables used as auxiliary information for the estimation of the weights – in the following called simply design variables – and the set of key variables used in the disclosure process. In Section 4.2 we focus our attention on the variability and bias of the risk estimates and present some results and comments for the entire set of 600 samples.

4.1 *Analysis of particular samples*

The first situation we investigate is the extreme situation when there is perfect agreement between the design variables and the key variables. This means that the two set of variables coincide completely and also their classifications coincide. In this extreme situation we know exactly the values of F_k and estimating the risk is not necessary. However, this extreme situation can be considered as a benchmark. Figure 1 shows the comparison between the estimated individual risk and the real risk in this situation for two samples among the 600; we obtained very similar results in other inspected samples. The plot is on the logarithmic scale. The set of common variables in Figure 1 is: age in 14 classes, sex, region of residence, position in profession and relationship with the head of household. To reach this situation the weights have been modified according to the procedure described in Section 3.2. In Figure 1 we distinguish plotted points according to their frequency in the sample. Unique cases, double cases, triplets and 4ples are plotted, respectively, by means of circles, triangles, pluses and crosses. The diamonds represent five or more cases.

To see how many combinations in the population correspond to these sample combinations it is sufficient to calculate the inverse of the value of *R*. Therefore, in both plots the highest value is a unique combination in the population and R =1; the second highest value is a combination that appears only twice in the population, and so on. For example, in the plot on the right there are two combinations in the sample that appear four times in the population; one has sample frequency equal to one and the other has sample frequency equal to two. The continuous line is the diagonal of the square [0,1] and represents the points where real and estimated risk are equal. Figure 1 shows a good agreement between the estimated and the real risk. Moreover, it is evident that the higher the frequency of the combination in the sample the better the agreement between the estimated and the real risk. However, sample uniques present a clear overestimation of the risk. This is due to the functional form of the estimator for sample uniques: $(\hat{F}_k - 1)^{-1} \log(1/\hat{F}_k)$. Moreover there is a clear pattern in overestimation. The higher the difference between the population frequency and its estimate, the higher the overestimation. This effect disappears for large values of $\hat{F}_k$.

To evaluate the impact of this overestimation we consider its consequences on the final result, i.e. the protection of the file. This means to take into account how many additional protection (e.g. local suppression) would be made because of the overestimation. In Figure 1 two examples of possible values of the threshold are plotted (the dotted line is at 0.04 and the dashed line is at 0.01). In both cases the additional combinations that would undergo protection due to overestimation are given by the points on the right of the vertical dotted (dashed) line and underneath the horizontal dotted (dashed) line, respectively. However, it has to be noticed that many of the sample uniques are due to unusual ages and therefore would undergo anyway a local suppression as all the microdata files that are released are treated with top coding of the age.

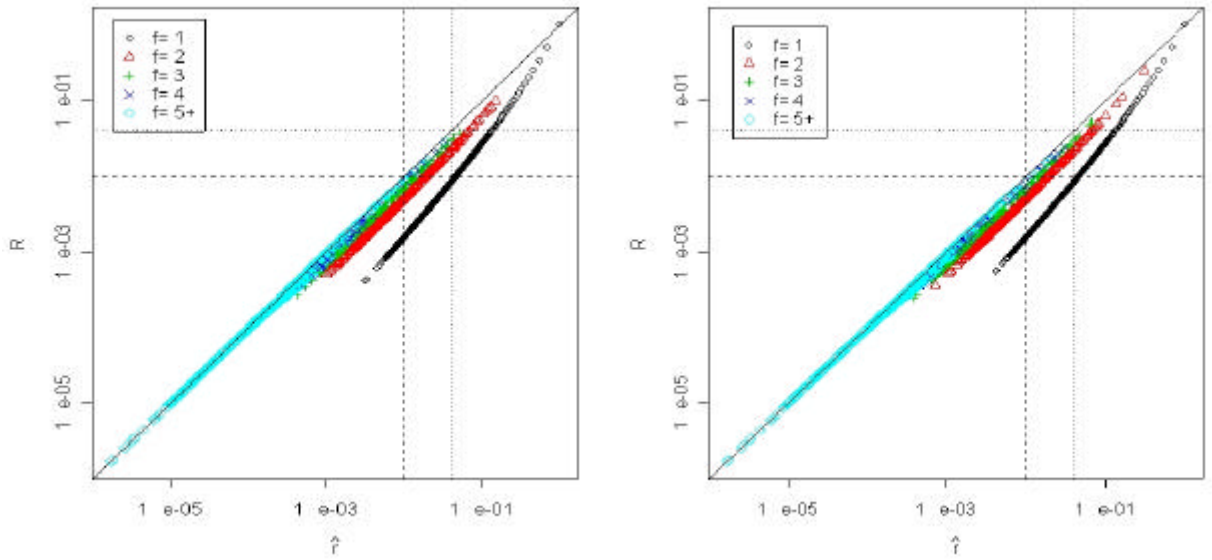
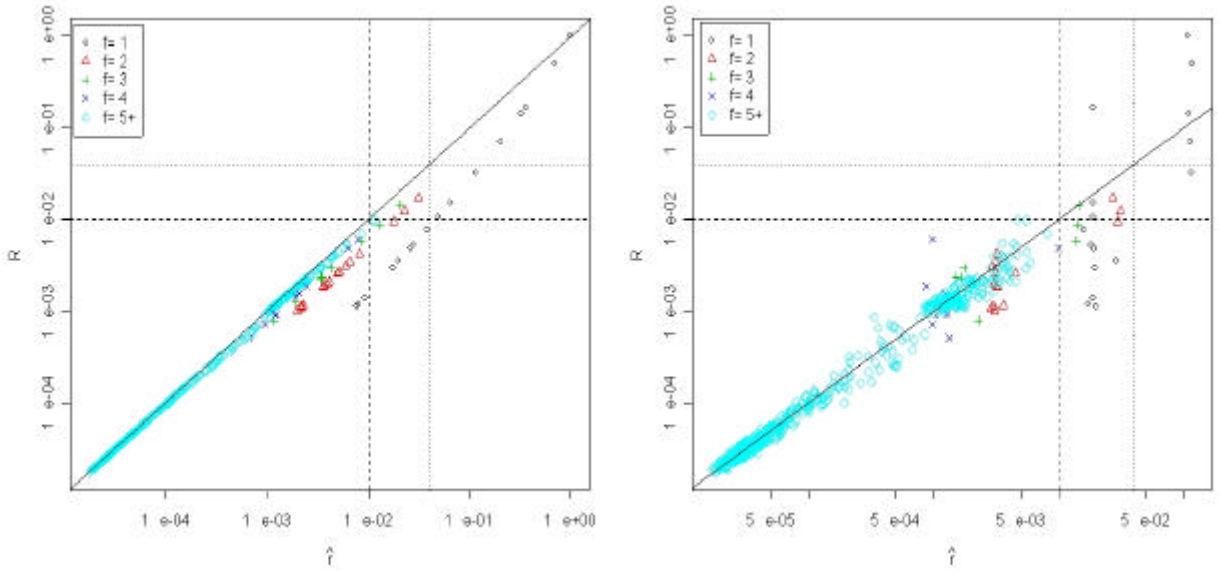


Figure 1: *Plot in the logarithmic scale of the real risk, R, vs. the estimated risk, $\hat{r}_i$, in two particular samples when there is perfect agreement between design and key variables (set of 5 variables). In the plot unique cases, double cases, triplets and 4ples are plotted, respectively, with circles, triangles, pluses and crosses. The diamonds represent five or more cases. The dashed and the dotted lines represent two different choices of the threshold value for subsequent application of local suppression.*



(a) (b)

Figure 2: *Plot in the logarithmic scale of the real risk, R, vs. the estimated risk, $\hat{r}_i$, for a particular sample when (a) there is perfect agreement between design and key variables: sex, region and age, and (b) sex and region remain unchanged but age is recoded in 14 classes in the design and in yearly classes in the key variable. In the plot unique cases, double cases, triplets and 4ples are plotted, respectively, with circles, triangles, pluses and crosses. The diamonds represent five or more cases. The dashed and the dotted lines represent two different choices of the threshold value for subsequent application of local suppression*.

Key variables	# combinations in the population	# combinations in the sample	# sample uniques (percentage)
Sex, region and age in yearly classes	859	768	25 (3.25)
Sex, region, age in 14 classes, profession and relationship with head of h/h	12526	2972	1173 (39.48)
Sex, region, age in yearly classes, profession and relationship with head of h/h	48981	7839	3637 (46.40)

Table 1: *Number of combinations in the population, in the sample and sample uniques for a particular sample when varying the set and classification of the key variables.*

Exactly the same pattern is reproduced for different sets of key variables. For example, Figure 2 (a) shows the case when the design and key variables are sex, region and age in 14 classes. Although the number of combinations is drastically reduced with respect to Figure 1, the pattern is clearly unchanged. To give an idea of the differences in the number of combinations in the population, in the sample as well as the number of sample uniques, we present in Table 1 these values for a particular sample among the 600 we created.
The conclusion therefore seems to be that that when there is perfect agreement between the design and key variables the unique cases are overestimated but in general there is good agreement on all the other cases.

However, in the real world, it is almost impossible to find perfect agreement between design and key variables. The investigation therefore moves towards the more realistic case where the design and key variables are the same but they have different classification. This is shown in Figure (2) (b) where the variable age in the design is used according to the 14 classes defined in Section 3.1 whereas in the key variable it consists of yearly classes. Figure 2 (b) shows underestimation as well as overestimation. The vertical pattern that is evident for sample rare combinations (frequencies between one and three) is due to the reduced variability of the weights that now present constant values inside the age classes. This means that, for example, all sample uniques assumes mainly two different weight values. This is not surprising if we notice that all of the sample uniques are due to unusual ages; recall that all ages greater than 75 belong to the same age band (indeed all the units sharing the same value of the risk have ages between 88 and 105) and this will clarify the pattern of this plot. Moreover, the large distance between these vertical lines is due to the sampling ratio differing according to the demographic size of the region.

Finally we consider the most common case where the design variables are a subset of the key variables. In particular we consider the case when the design variables are those described in Section 3.1 and the key variables present, in addition to those, also the position in the profession and the relationship with the head of the household. This situation is plotted in Figure 3 for a particular sample. As in Figure 2 (b) the vertical pattern is present in the plot. As in Figure 2 (b) overestimation as well as underestimation of the real risk is evident. The presence of underestimation of the risk implies that combinations that should undergo local suppression remain, in reality, untouched. This is the case for all the combinations that are above the horizontal dotted (dashed) line and on the left of the vertical dotted (dashed) line, respectively. As far as overestimation is concerned, for the threshold value equal to 0.01, the total number of units that present a risk higher than the threshold is equal to 1422; of these 636 (45%) present a real risk that is lower than the threshold (for the other threshold the value, 0.04, they are, respectively, 222 and 82 (37%)).

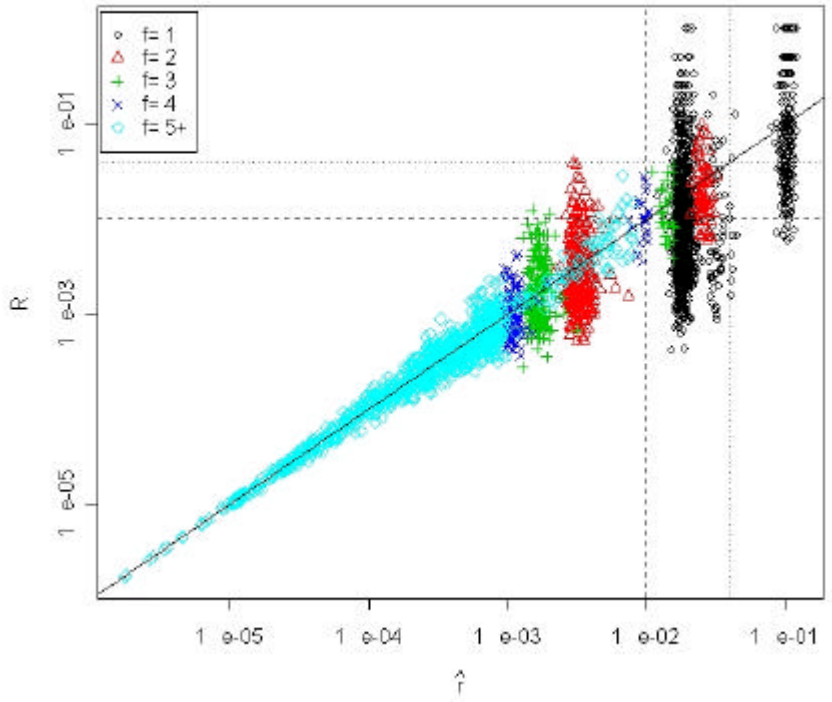


Figure 3: *Plot in the logarithmic scale of the real risk, R, vs. the estimated risk, $\hat{r}_i$, in a particular sample when the design variables are sex age in classes and region and the key variables are sex, age in classes, region, position in profession and relation with the head of the household. In the plot unique cases, double cases, triplets and 4ples are plotted, respectively, with circles, triangles, pluses and crosses. The diamonds represent five or more cases. The dashed and the dotted lines represent two different choice of the threshold value for subsequent application of local suppression.*

A note on the behaviour of the estimates of the risk when small regions are involved. In Figure 3 the vertical band presenting the highest risk in the sample is all from combinations stemming from Valle d'Aosta, the smallest region of the four examined. This because such region presents the highest sampling rate; this is so in order to obtain errors that are of the same order as the other Italian regions. Larger sampling rate implies smaller weights and therefore higher individual risk.

4.2 *Analysis of the whole set of samples*

To investigate the variability of the individual risk estimates we selected a set of combinations that are present in all 600 samples. They are all the women with residence in Lazio and age between 0 and 83. The boxplot of the individual risk estimates is shown in Figure 4; the box contains 50% of the observations and the median of the distribution is indicated by the horizontal line in the box. The wiskers are positioned at 1,5 of the interquartile range; outliers are indicated with a circle. In order to make the comparison, the value of the real risk is plotted as well; this is the large filled dot.

The estimated individual risk seems to have a good behaviour with respect to the real value of the risk. In all cases the real risk either is equal to the median or it is very near to it. Notice that the boxplots show the same variability for the ages that in the design are in five year classes whereas they show higher variability for those ages that are in the first, [0,14], and last, [75,-], classes. This seems to validate the hypothesis that a complex design with a calibrator estimator for the weights is a good starting point for the estimate of the risk. However, we will further investigate this. It is interesting to note that the sudden big increase in the real risk between 71 and 72 years (there are 26626 women in Lazio who are 71 years old but only 15408 are 72 years old) is well captured by the estimated risk.

We replicated the same plot also for ages between [0,90]. This was to investigate variability for less common combinations. For example in this set of combinations the minimum frequency in the sample is 1 and the minimum in the population is 2651; the maximum in the sample is 186 and the maximum in the population is 44,943.

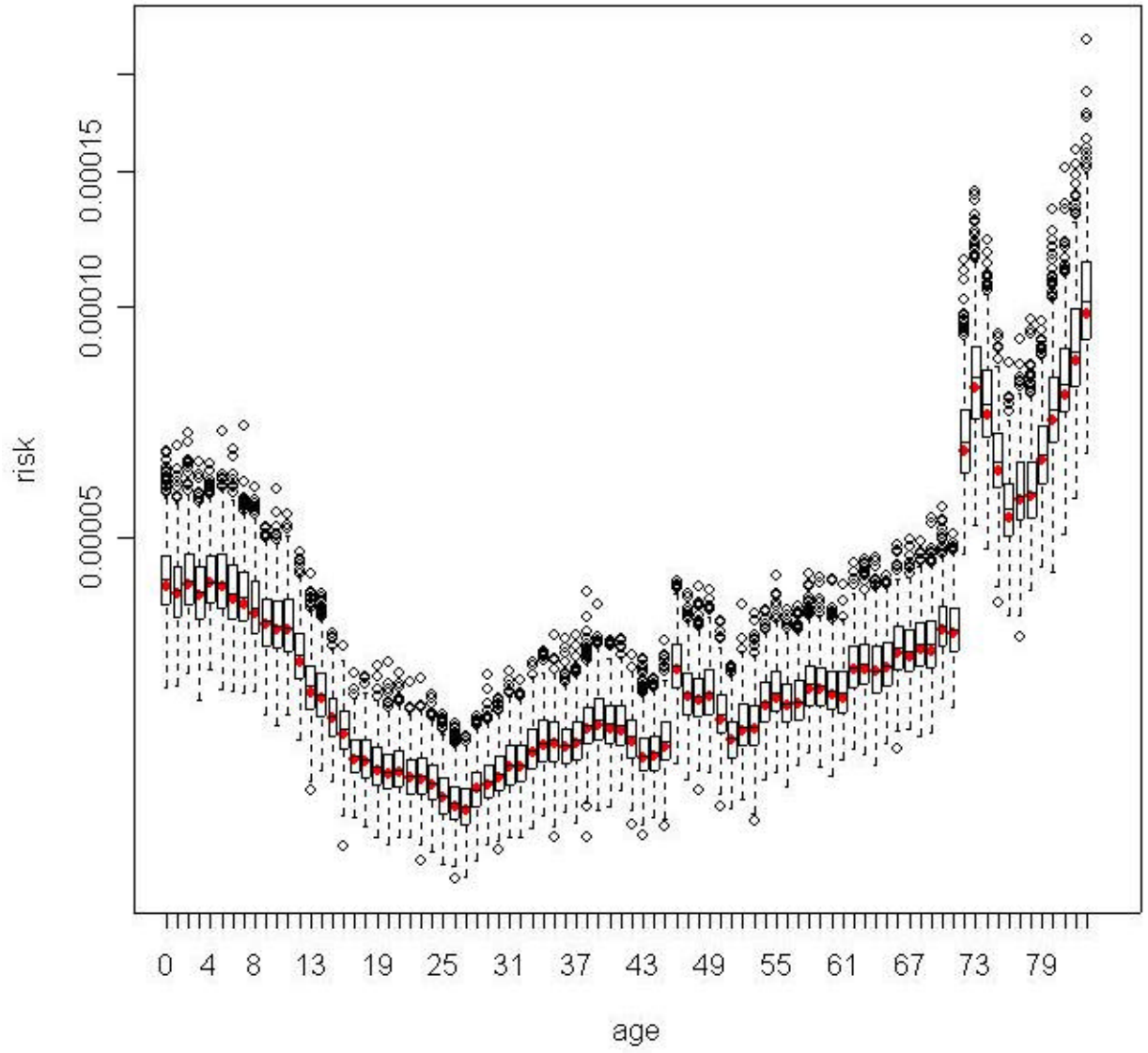


Figure 4: *Boxplot of the log of the individual risk estimates, $\hat{r}_i$, for each age yearly class between 0 and 83, for the women with residence in Lazio when the key variables are sex, region and age in year and the design variable age is in 14 classes. The real risk, R, is represented by the filled dot. All these combinations appear in all the 600 samples.*

5. Conclusions and suggestions for further work

We present the results of an experiment for assessing the performance of the individual risk model proposed by Benedetti *et al.* (1999) and currently under implementation in the software μ-Argus as part

of the European project CASC. On the basis of the 1991 Italian population census for four regions we created 600 samples using the LFS sampling design.

We analyse the behaviour of the individual risk estimates in the realistic case of departure from the optimal condition of complete correspondence between design and key variables. We noticed a good agreement between the real risk in the population and the estimates of the individual risk although the quality of the estimator seems poor for rare combinations in the sample as compared to the more common ones. Indeed, rare occurrences in the sample (like sample uniques) can stem from both rare and common features in the population and this just because of the sampling structure. 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.

Further experiments are needed to assess the behaviour of the estimates of the individual risk for sample uniques arising from rare combinations in the population. Indeed six hundreds samples are not enough to investigate rare occurrences in the population. Moreover, we intend to examine also other sets of key variables to see how much the departure from the complete agreement between the design and key variables is influencing the risk and what consequences this has on the protection of the microdata file. Finally, further studies to improve the performance of the estimator used for the individual risk have been planned; in particular, it is foreseen the use of methodologies borrowed from the area of model assisted small domain estimation.

Acknowledgments

The authors are very grateful to Silvia Polettini for comments and help with the graphs. In addition we would like to thank Yosi Rinott for suggesting the conduction of an experiment to assess the individual risk model.

This work was partially supported by European project IST-2000-25069 on “Computational Aspects of Statistical Confidentiality”.

The views expressed are those of the authors and do not represent the policy of Istat.

References

Benedetti, R. and Franconi, L. (1998), Statistical and technological solutions for controlled data dissemination, Pre-proceedings of New Techniques and Technologies for Statistics–Sorrento, 4-6 November 1998, vol.1, 225-232.

Benedetti, R., Franconi, L. and Piersimoni, F. (1999), Per-record risk of disclosure in dependent data, Proceedings of the Conference on Statistical Data Protection, Lisbon 25-27 March 1998. European Communities, Luxembourg.

Cochran, W. G., (1977), Sampling Techniques, Wiley, New York.

Deville, J. C., Särndal, C. E., (1992), Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, 87, pp. 367-382.

Duncan, G.T. and Lambert, D. (1986), Disclosure limited data dissemination, (with discussion), Journal of the American Statistical Association, 81, 393, 10-28.

Fienberg, S.E. and Makov, U.E. (2001), Uniqueness, urn models and disclosure risk. *Research in Official Statistics*, 17, 499-520.

Madow, W.G. (1949), On the theory of systematic sampling II, The Annals of Mathematical Statistics, 20, pp333-354.

Polettini, S. (2003), Some remarks on the individual risk methodology, To be presented at the Joint ECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003).

Skinner, C.J. and Elliot, M.J. (2002), A measure of disclosure risk for microdata, Journal of the Royal Statistical Society, Series B, 64, 855-867.