Individual Risk of Disclosure Using Sampling Design Information

R. Benedetti * A. Capobianchi † L. Franconi †

Abstract

National Statistical Institutes routinely release Microdata File for Research i.e. individual data which have been modified in order to minimise possible disclosure of confidential information. An assessment of the risk of disclosure is always advisable before releasing the data. In this paper we show the importance and flexibility of an individual risk of disclosure in contrast with an aggregated measure of the risk for the whole microdata file. Moreover, when discrete key variables are present, we propose a new method for estimating the risk of each single unit in the file to be released quantifying the probability of correctly linking each unit to an individual in the population. We formalise the relationship between frequencies in the sample and those in the population by mean of sampling weights, showing the impact that survey design has on disclosure limitation. An assessment of the proposed method is provided by comparing our estimates of the individual risk with those obtained using demographic information. The results of such a comparison based on the Italian Household Expenditure Survey are presented.

Key words: Statistical disclosure limitation; Microdata File for Research; Discrete key variables; Negative Binomial Distribution.

1 Introduction

The growing demand by the scientific community to allow statistical analysis to be made on individual data has fostered research into the field of statistical

*Dipartimento di Economia, Università G.d'Annunzio, Chieti, Italy

†Istituto Nazionale di Statistica, MPS, Via C. Balbo 16, 00184, Roma, Italy. E-mail: franconi@istat.it

disclosure limitation (Doyle *et al.*, 2001, and Domingo-Ferrer, 2002). It has also led recently to a crucial change in European legislation with the adoption of Commission Regulation no. 831/2002 on "access to confidential data for scientific purposes". This law would allow the creation of "anonymised microdata", essentially what is known as Microdata File for Research (MFR), at European level for some strategic surveys.

Any microdata file to be released will contain a set of variables for each unit i: the *key variables* that may allow identification of an individual and are accessible to the public, the *sensitive variables* that are confidential and the *final weight* w_i indicating the number of individuals in the population that are represented by unit i (Deville and Särndall, 1992). We focus our attention on discrete key variables; this is the common situation in household surveys and population censuses.

This paper is motivated by the production of MFR stemming from social surveys (the file to be released is, therefore, a sample from the population of interest). This involves both the development of techniques to protect the data and the definition (and evaluation) of a measure to quantify the probability that a user has to disclose individual information; this paper concentrates on the latter. The individual disclosure risk we provide uses sampling design information. It is therefore particularly suited for social surveys carried out by National Statistical Institutes (NSIs) that put extreme care in the design of their surveys. Our aim in this paper is to present a methodology that can be used by NSIs to routinely produce MFRs. To this end such methodology has been included in the software μ-Argus, a software package for producing safe microdata files, discussed by Willenborg and Hundepool (1999). Currently the software, a product of the 'Computational Aspect of Statistical Confidentiality' project founded by the European Union is under testing for its final release at the end of 2003. More information on the project is available from the Web page `http://neon.vb.cbs.nl/casc/`

μ-Argus includes both the risk assessment and the protection of microdata. To do this it allows to estimate the risk as described in this paper for each unit in the microdata file to be released. Moreover, it applies global recoding in an interactive way and, for all those units presenting a risk higher than a predefined threshold, it performs local suppression in order to produce a safe microdata file.

In this paper we adopt the re-identification disclosure definition (Willenborg and de Waal, 2001). This occurs when an individual unit is identified (i.e. a one-to-one relationship between a unit in the released file and a target individual in the population is established with some degree of confidence) and then, as a consequence, the user is able to deduce the value of sensitive variables for such individual. In Section 1.1 we will specify the

way in which an identification can be performed.

1.1 Disclosure scenario

To develop a quantitative model for measuring individual risk of disclosure we need first of all to make some assumptions on the behaviour of the person who attempts the identification, therein called the *intruder*, and also define what we mean with the term *identification*.

As far as the intruder behaviour is concerned, we assume that:

-the intruder has an external data base or public register available that contains identifiers (for example names and addresses of individuals) and key variables;

-no measurement error in the value of the key variables occurred;

-the intruder tries to identify the unit i in the sample comparing the observed combination of categories of the key variables for such unit in the released sample with the same combinations for individuals i^* on his/her register.

As far as the term identification is concerned, we observe an identification of a unit i in the released sample when two conditions are met. First this unit is linked, through the values observed on the key variables, to an individual i^* contained in the register available to the intruder and, second, i^* is the individual of the population from which the unit i is derived. If only one individual in the register and one unit in the released sample present the same categories of the key variables a one-to-one relationship occurs. In this case if the external data base covers the whole population the intruder identifies the individual. For those individuals for whom it is not possible to find a one to one relationship between the categories of the key variables in the two data-sets, the link will be based on probabilistic reasoning as it will be discussed in Section 3. As a consequence, even if the external register covers the whole population, the intruder is not certain of identifying the individual. The previous remark shows how the dimension of the external register is crucial. The worst situation for the NSI releasing the sample occurs when such external register covers the whole population. In fact, only in this case the intruder can be sure of identifying some individuals (the one-to-one relationships) and, in general, he has the highest chance of observing an identification. Therefore, to consider the worst situation, in what follow we assume that the external register covers the whole population.

1.2 Reasons for an individual risk of disclosure and outline of the paper

Broadly speaking there are two different approaches to microdata risk assessment: a global approach and an individual approach. The former provides a single figure for the whole microdata file to be released whereas the latter produces a risk for some (or all) units in the file to be released. The global risk approach mainly concentrates on unique cases i.e. units presenting unique combination of scores on the key variables. The use of these frequencies to assess the risk of disclosure is a common practice in many NSIs. This practice is based on the assumption that only population uniques can be identified and on the evidence that a population unique is necessarily a sample unique. However, as the contrary is not always true, this approach may lead to an over protection of the microdata file. In general the global approach develops models for estimating the expected number of population uniques given the sample uniques. Recent examples of the use of this approach can be found in Fienberg and Makov (2001), Hoshino (2001), Samuels (1998) and references therein.

From the point of view of information loss, a global approach leads to the application of protection techniques that are *variable driven*, such as global recoding: Willenborg and de Waal, (2001). An individual approach, on the contrary, allows for *unit driven* protection methods, such as local suppression: Willenborg and de Waal, (2001). Unit driven methods, being selective i.e. applying only to those units presenting a risk higher than a predefined threshold, result in limited information loss.

From the point of view of the level of safety in the released file the limitation of the global approach stands in treating all units presenting the same frequency with respect to the key variables as exchangeable thus ignoring different combinations of scores on those variables. However, such frequencies are not all alike; Fienberg and Makov (2001), in a contingency table view of the problem, recall how these may correspond to cells with very different underlying probabilities. The individual risk approach tries to find ways to include these differences into the disclosure risk. Skinner and Holmes (1998) and Fienberg and Makov (1998) express such differences in terms of log-linear models for contingency tables. Skinner and Elliot (2001) propose a new measure of disclosure risk as the probability that a unique match between a microdata record and a population unit is correct. In all these approaches, a risk is estimated only for units presenting unique combinations of score on the key variables. In this paper we provide a risk of disclosure for each unit in the microdata file to be released using sampling design information.

In Section 2 we outline the definition of an individual disclosure risk and

in Section 3 we present an estimation method of the proposed risk based on sampling information. In Section 4 we appraise the new method on data from the Italian Household Consumption survey. We end the paper by briefly presenting our conclusion in Section 5.

2 Individual Disclosure Risk

Let s be the observed sample of size n selected from a finite population of N individuals according to a design D. Our aim is to release this sample protecting the confidentiality of individual respondents. To reach this aim we define a model to assess the possibility of disclosure of confidential information for each unit in this sample and allow the protection of those units presenting a risk higher than a predefined threshold.

For each unit i we define the disclosure risk, r_i, as the probability of identifying such unit given the information contained in the observed sample. This is the probability of linking unit i in the sample to individual i^* in the register given the observed sample:

$$r_i = \Pr(\text{unit } i \text{ is linked to the individual } i^* \mid s)$$

where i^* is the individual from whom the unit i is derived.

In order to simplify notation, in what follow let L_i be the event "unit i is linked to individual i^* where i^* is the individual from whom the unit i derived", so we have: $r_i = Pr(L_i \mid s)$.

Note that the disclosure risk is defined only for units in the sample to be released as an identification can not take place otherwise. Moreover, units in the sample who share the same combination of categories of the key variables are identical for the intruder in terms of uncertainty to make an identification.

Hence we can restrict the analysis on each of the $k = 1, \ldots, K$ domains defined by all the possible combinations of categories of key variables. For a particular unit i in the sample to be released let $k(i)$ be the domain it belongs to; then the risk of each unit in the domain $k(i)$ is equal to the risk of the unit i. In what follow, we denote with $r_{k(i)}$ the risk of each unit in the domain $k(i)$, and $r_{k(i)} = r_i$. The aim is, thus, to estimate the quantity r_i for each unit i in the sample.

3 Estimating the Individual Disclosure Risk

Let f_k and F_k be, respectively, the number of units in the released sample and the number of individuals in the population in the k-th domain; F_k is

unknown for each k. In the sample to be released only a subset of the total number K of domains will be observed and only this subset, for whom $f_k > 0$, is of interest to the disclosure risk estimation problem.

The information given by the sample for the identification of unit i consists of the frequency $f_{k(i)}$. Hence we can write r_i as:

$$r_i = r_{k(i)} = P(L_i \mid s) = P(L_i \mid f_{k(i)}).$$

Such conditional probability can be express as:

$$\begin{aligned} r_{k(i)} = \Pr(L_i \mid f_{k(i)}) &= \Pr(L_i \mid F_{k(i)} = 1, f_{k(i)}) \Pr(F_{k(i)} = 1 \mid f_{k(i)}) + \\ &\quad \Pr(L_i \mid F_{k(i)} = 2, f_{k(i)}) \Pr(F_{k(i)} = 2 \mid f_{k(i)}) + \ldots \\ &= \sum_{h \geq f_{k(i)}} \Pr(L_i \mid F_{k(i)} = h, f_{k(i)}) \Pr(F_{k(i)} = h \mid f_{k(i)}) \end{aligned} \quad (1)$$

other terms in the summation disappearing because of the assumption of no measurement errors which implies $F_{k(i)} \geq f_{k(i)}$. To evaluate (1) some distributional assumptions on the unknown parameter $F_{k(i)}$ given the observed frequency $f_{k(i)}$ are necessary. A possible solution is to estimate a loglinear model for the multi-way table $f_{k(i)}$ and then compute the probability of $F_{k(i)}$ either directly, see Skinner and Holmes (1998), or using the estimated model for the imputation of nonsampled individuals, see Fienberg and Makov (1998). The former approach is attractively distributional free, although our experience is that its practical application is made difficult by the time required and the loss of numerical accuracy when estimating models with more than 1000 domains (10% of which has structural or sampling frequency equal to zero). To avoid such problems we use the sampling design to perform a direct estimation of the risk. Let us consider $\Pr(L_i \mid F_{k(i)} = h, f_{k(i)})$. When the external register covers the whole population, if unit i in the sample is unique in the population then $\Pr(L_i \mid F_{k(i)} = 1, f_{(k(i)}) = 1$. If two individuals $i^\star$ and i' belong to the same domain $k(i)$ then the intruder performing a probabilistic linkage obtains $\Pr(L_i \mid F_{k(i)} = 2, f_{(k(i)}) = \frac{1}{2}$. If the same probabilistic reasoning is applied to higher values of $F_{k(i)}$, then equation (1) can be written as:

$$r_{k(i)} = \sum_{h \geq f_{k(i)}} \frac{1}{h} \Pr(F_{k(i)} = h \mid f_{k(i)}). \quad (2)$$

In order to simplify the notation in what follows we omit the subscript i.

As far as the factor $\Pr(F_k = h \mid f_k)$ is concerned the idea is to consider $F_k \mid f_k$ as a random variable distributed according to a negative binomial

distribution with f_k successes and probability of success p_k, see Skinner *et al.* (1994) and Bethlehem *et al.* (1989). Then,

$$\Pr(F_k = h \mid f_k) = \left(\begin{array}{c} h-1 \\ f_k - 1 \end{array} \right) p_k^{f_k} (1-p_k)^{h-f_k} \qquad h = 1, 2, \ldots$$

where $h \geq f_h$ and $f_k > 0$. This superpopulation approach is based on the assumption that having observed f_k *successes*, where a success is a selection with probability p_k of an individual from F_k, the allocation scheme can be described by an inverse binomial sampling. Therefore in equation (2) we recognise the negative moment of order 1 of the negative binomial distribution: $r_k = \mathrm{E}(F_k^{-1} \mid f_k)$. From Cressie *et al.* (1981), $\mathrm{E}(F_k^{-1} \mid f_k) = \int_0^\infty M_{F_k|f_k}(-t)dt$ where $M_{F_k|f_k}$ is the moment generating function of $F_k \mid f_k$. The substitution of the moment generating function of the negative binomial distribution leads to:

$$r_k = \int_0^\infty \left\{ \frac{p_k \exp(-t)}{1 - q_k \exp(-t)} \right\}^{f_k} dt \tag{3}$$

where $q_k = 1 - p_k$. The transformation $y = \{1 - q_k \exp(-t)\}^{-1}$ in the integral (3) gives:

$$r_k = \left(\frac{p_k}{q_k} \right)^{f_k} \int_1^{\frac{1}{p_k}} \frac{(y-1)^{f_k - 1}}{y} dy \tag{4}$$

which is a monotonically increasing function in p_k and monotonically decreasing in f_k and F_k; see Figure 1 and Figure 2.

To estimate the risk we need to estimate the parameter p_k for each domain k. Our idea is to make use of the sampling framework commonly employed by NSIs to make inference from samples to populations. In fact, to get accurate and efficient estimates of the phenomenon under study NSIs adopt complex sampling design for their surveys, for example multistage and stratified. We believe that such an effort can be also useful for the purpose of estimating the individual disclosure risk. To do this, in the maximum likelihood superpopulation estimator of p_k, $\hat{p}_k^{MLE} = \frac{f_k}{F_k}$, we introduce the information given by the sample through the design D. Such information can be summarised by the Horvitz-Thompson estimator of F_k:

$$\widehat{F}_k = \sum_{i:\ k(i)=k} w_i \tag{5}$$

where w_i is the final sampling weight of unit i in domain k. Finally, we obtain the *design based* estimator of p_k as:

$$\widehat{p}_k^{Des} = \frac{f_k}{\sum_{i:\ k(i)=k} w_i}. \tag{6}$$

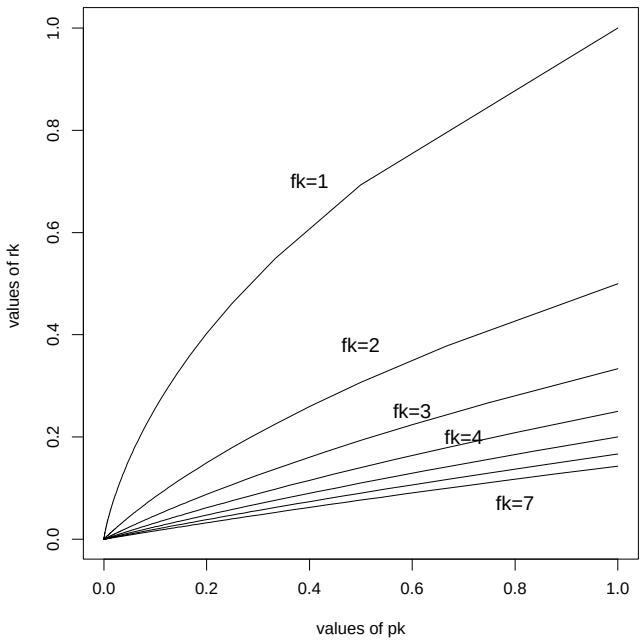

Figure 1: Risk of disclosure as function of the probability of selection p_k for different values of the observed frequencies f_k.

3.1 Factors Influencing the Risk

The methodology we propose can accommodate factors influencing the risk and that do not depend on released data. This is true, for example, for the degree of coverage of the register, for the quality of the data as a function of the time lag between the conduction of the survey and the data release. In particular, given that these additional factors relate to events that are independent from each other, we can formalise the probability of identification of individual i as the product of the probabilities of each factor. For example, let $d_{k(i)}$ be a quality parameter, $e_{k(i)}$ be the probability of the unit to be included in the external file (which depends usually on the availability of public registers and on their degree of coverage of the population) and $t_{k(i)}$ be the probability of an attempt of identifying unit i.

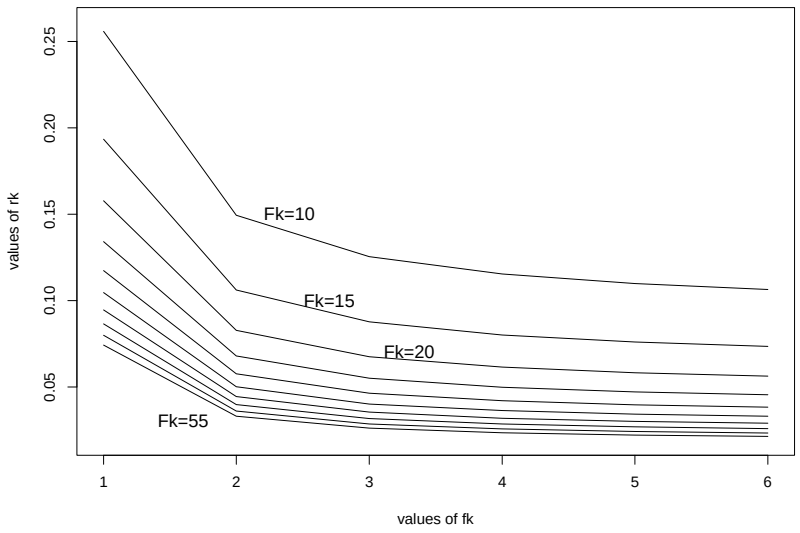

Figure 2: Risk of disclosure as function of the observed frequencies f_k for different sizes of subpopulation F_k.

Then, the final risk can be written as:

$$\rho_{k(i)} = d_{k(i)} e_{k(i)} t_{k(i)} \hat{r}_k,$$

where $\hat{r}_k$ is the value of r_k when estimated through the use of $\widehat{p}_k^{Des}$. Of course each of these parameters can vary across units or groups of units. As a final remark, notice that in this formalisation $\hat{r}_k$ is the only parameter depending on released data and any additional information has to be known *a priori*. This approach can be thus considered in a Bayesian framework (Fienberg, Makov and Sanil, 1997, Duncan and Lambert, 1986) where $\hat{r}_k$ plays the role of the likelihood.

4 Case Study Based on Italia Household Consumption Survey

In this section we present the application of the proposed methodology to the Italian Household Consumption Survey (HCS) relative to the year 1997 in order to assess the methodology main characteristics.

The HCS is a survey based on a two stage sample design in which the primary sampling units are the municipalities and the secondary

sampling units are the households. The municipalities are stratified by size (population) and NUTS3 (Nomenclature of Statistical Territorial Units) geographical code (provinces). The households are selected independently without replacement and with equal probability from the sampled municipalities. The survey provides household consumption estimates significant at regional (NUTS2) level.

The survey data analysed comprises 64,000 units (22,363 households).

The final sampling weights w_i are evaluated using a calibration estimator (see Devil and Särndal, 1992). They are calculated solving a minimum constraint problem where the constraints are defined by the equality between the survey estimates of some population totals and the corresponding known population quantities. The HCS constraints are relative to the known distribution of sex by age in each region, where the age is recoded in four classes (0-14, 15-29, 30-59, 60-100).

To assess the design based estimation approach we compare the values of the estimates of the individual risk using the design based estimator (6), with estimates of the individual risk using demographic information: $\widehat{p}_k^{Dem} = f_k/\widehat{F}_k^{Dem}$, where $\widehat{F}_k^{Dem}$ is the value of the Italian population relative to the k-th domain as it is calculated by Istat in the context of demographic analysis of Italian Population (Istat 2000). Such estimates are calculated using administrative sources. This comparison investigates whether sampling weights, designed to estimate quantities inherent to a particular survey, can be adopted to solve the individual risk estimation problem.

For this comparison we use two overlapping sets of key variables. The first set contains sex (2 categories), age (99 categories) and region (20 categories) while the second set comprises also the variable marital status (6 categories). Notice that all the key variables in the first set appear as constraints in the calibration estimator whereas, in the second set, marital status is not present in the design estimator constraints.

In Figure 3 we plot, for the first set of key variables, the values of the risk obtained by the design based estimates against the values of the risk evaluated through demographic estimates. Units lying on the diagonal line share the same value of the risk. We see that there is a high correlation between the two sets of risk values. As expected, large values of the risk show higher variability than small values.

As already mentioned before, units showing a value of the risk higher than a predefined threshold α undergo a protection technique (global recoding, local suppression etc.). In Figure 3 we plot a possible value for such a threshold. Taking into account the values of the factors influencing the risk for this survey commonly used at ISTAT, such threshold corresponds to the probability of identification a unit equal to one over 40000.

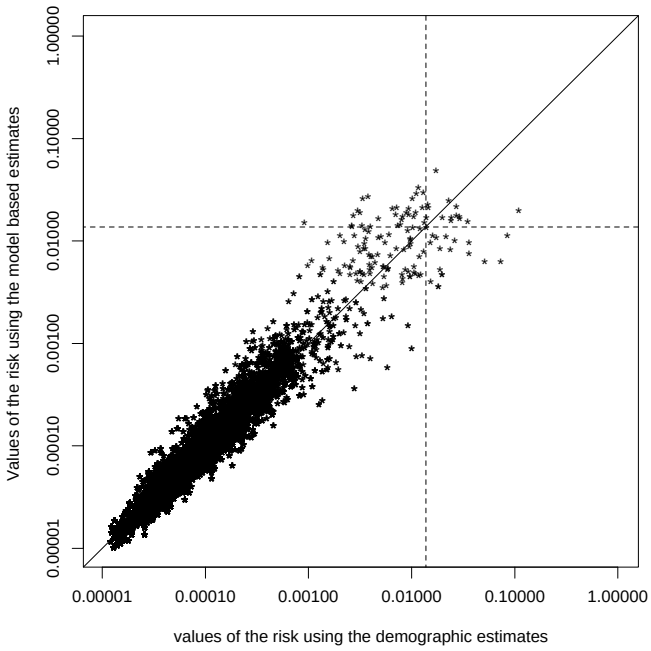

Figure 3: Risk using model based estimates and the demographic estimates on key variables sex, age and region; HCS data. A logarithmic scale is used. Units lying on the diagonal line share the same value of the risk. The dashed lines indicate a possible value for the threshold.

As it is evident from Figure 3 there is no bias in the behaviour of the model based estimates. However, two types of errors can occur: the first involves information loss, the second the level of safety of the released data. In fact, when a design based estimate of the risk is above the threshold but the corresponding demographic estimate isn't, then we may overprotect the data to be released applying local suppression to the corresponding unit. On the other hand, if the design based estimate of the risk is below the threshold and the corresponding demographic risk is above we will not protect a unit that may be at risk. These instances are, however, extremely rare. The first type of error occurs for 30 units (0.04 % of the total), whereas the second type of error occurs for 18 units (0.028% of the total). We conclude that

the design based method estimates reasonably well the individual risk of disclosure when the key variables coincide with the constraint variables in the calibration estimator.

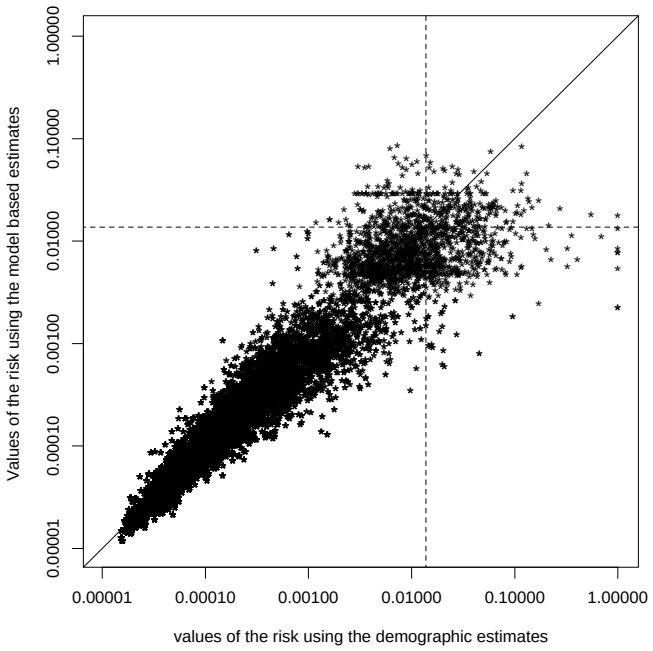

Figure 4: Risk using model based estimates and the demographic estimates on key variables sex, age, region and marital status; HCS data. A logarithmic scale is used. Units lying on the diagonal line share the same value of the risk. The dashed lines indicate a possible value for the threshold.

In Figure 4 we plot the values of the risk based on the second set of key variables. As in Figure 3 there is still high correlation between the two sets of risk values, however higher variability than the previous case is shown. A reason for this is that the new key variable introduced, marital status, has not been included neither in the planning of the sampling design nor in the definition of the calibration estimators. In particular, we observe that the first type of error, as defined above, occurs in 190 units (0.29% of the total),

whereas the second type of error occurs in 404 (0.63% of the total) with an increase of 0.24% and 0.57% respectively.

From the comparison of Figure 3 and Figure 4 and the considerations made above, we argue that the design based risk estimation is a valid methodology to evaluate the individual risk of disclosure. This is true also in cases when there is no perfect coincidence between the set of variable used to define the sampling design and the set of the key variables used to define the disclosure problem.

5 Conclusion

Very frequently the decision to release a MFR relies on measures of disclosure risk based on the frequency of occurrences of sample uniques. In this paper we show the limitation of such measure and argue on the need to relax the hypothesis of exchangeability underneath this approach. To overcome such limitations we outline an individual risk of disclosure based on negative binomial distribution. This method provides a theoretical answer to the need for risk estimation methods able to exploit all the available information underlying the survey, with particular reference to the sampling design. The proposed estimation method performs a direct extension to the population by means of sampling weights and requires a very small computational burden. In its application to the Italian HCS, this methodological approach has proved to provide valid estimates of the individual risk even in those cases where the set of the sampling design variables did not coincide with the set of key variables.

Finally this individual approach is essential to formalise, in a flexible way, the introduction of factors not depending on released data which may influence the disclosure risk. This is the case of the quality of the key variables, the size of the intruder data-base and any other source of noise which can be different from unit to unit.

The flexibility, the validity and the simple implementation and use through the software μ-Argus show that such method has the potentiality to solve the problems of practical survey data protection.

Acknowledgements

The authors are very grateful to Dr Julian Stander for useful suggestions on previous versions of this paper. In addition we would like to thank Giuliana Coccia for providing the data and Giovanni Seri and Silvia Polettini for

helpful comments.

Alessandra Capobianchi and Luisa Franconi gratefully acknowledge the financial support of the European Union, contract IST-2000-25069, CASC project.

The views expressed are those of the authors and are not intended to represent the policies of Istat.

References

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1989). Disclosure control of microdata. *Journal of the American Statistical Association,* **85**, 38–45.

Cressie, N., Davis, A. S., Folks, J. L. and Policello, G. E. (1981). The moment-generating function and negative integer moments. *American Statistician,* **35**, 148–150.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* **87**, 376–382.

Domingo-Ferrer, J. (2002). *Inference Control in Statistical Databases, Lecture Notes in Artificial Intelligence.* Springer-Verlag.

Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, L. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* Elsevier Science.

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. (1998). Confidentiality, uniqueness and disclosure limitation in categorical data. *Journal of Official Statistics,* **14**, 4, 385–397.

Fienberg, S.E., Makov, U.E. and Sanil, A.P. (1997). A Bayesian approach to data disclosure: optimal intruder behavior for continuous data. *Journal of Official Statistics,* **13**, 1, 75–89.

Fienberg, S.E. and Makov, U.E. (2001). Uniqueness, urn models and disclosure risk. *Research in Official Statistics*, **1**, 23–40.

Hoshino, N. (2001). Applayng Pitman's Sampling Formula to Microdata Disclosure Risk Assesment. *Journal of Official Statistics*, **17**, 4, 499–520.

Istat (2000). Popolazione per sesso, età e stato civile nelle province e nei grandi comuni. Nuove stime per gli anni 1993-1998. Sistema statistico nazionale. Istituto nazionale di statistica. Roma.

Lambert, D. (1993). Measure of disclosure risk and harm. *Journal of Official Statistics,* **9**, 313–331.

Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics,* **14**, 361–372.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics,* **10**, 31–51.

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.

Skinner, C.J. and Elliot, M.J. (2001). A Measure of Disclosure Risk for Microdata. Paper submitted for pubblication.

Willenborg, L. and Hundepool, A. (1999). ARGUS: software from the SDC project. *Statistical Data Confidentiality: Proceeding of the Joint Eurostat/ UN-ECE Work Session on Statistical Data Confidentiality*, March 8–10, 1999, Thessaloniki, 87–98.

Willenborg, L. and de Waal, T. (2001). *Elements of statistical disclosure control*, Springer-Verlag, New York.