WP. 8
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 UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (ii): Disclosure risk, information loss and usability of data

ASSESSING RISK IN STATISTICAL DISCLOSURE LIMITATION

Invited Paper

Submitted by the University of Naples, Italy and
the University of Plymouth, United Kingdom[1]

[1] Prepared by Silvia Polettini, University of Naples, Italy (spolettini@unina.it) and Julian Stander, School of Mathematics and Statistics, University of Plymouth, U.K. (jstander@plymouth.ac.uk).

Assessing Risk in Statistical Disclosure Limitation

Silvia Polettini*, Julian Stander**

* University of Naples, Italy. (spolettini@unina.it)

** School of Mathematics and Statistics, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK. (jstander@plymouth.ac.uk)

Abstract. When microdata files for research are released, it is possible that external users may attempt to breach confidentiality. For this reason most National Statistical Institutes apply some form of disclosure risk assessment and data protection. Risk assessment first requires a measure of disclosure risk to be defined. In this paper we follow the superpopulation approach of previous work and discuss a variety of Bayesian hierarchical models for risk estimation. For each combination of values of the key variables we obtain an estimate of disclosure risk. We discuss the models in detail and compare their performance by using an artificial sample of the Italian 1991 Census data, drawn by means of a widely used sampling scheme. Finally, we discuss further models for assessing risk in statistical disclosure limitation.

1 Introduction

When microdata files for research are released, it is possible that external users may attempt to breach confidentiality. For this reason most National Statistical Institutes apply some form of disclosure risk assessment and data protection. Risk assessment first requires a measure of disclosure risk to be defined; as this is usually cast in terms of population quantities, risk estimation is then achieved by introducing suitable statistical models. If the estimated risk is considered not tolerable, protection measures must be put into practice.

We base our definition of disclosure on the concept of re-identification; see Willenborg and de Waal (2001). Therefore by disclosure we mean 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.

Even when attention is focused only on re-identification disclosure, different approaches to risk assessment can be pursued. For instance, global risk measures can be defined that allow us to screen out unsafe data releases; see, for example, Fienberg and Makov (1998), Duncan and Lambert (1989), Bethlehem, Keller and

Pannekoek (1990), Lambert (1993), Skinner and Elliot (2002), and Carlson (2002). Alternatively, individual or combination-level risk measures, as defined in Benedetti and Franconi (1998), Skinner and Holmes (1989), Carlson (2002), and Elamir and Skinner (2004) among others, can be exploited to identify and protect unsafe records before the microdata file is released. A routine for computing a measure of individual risk of disclosure is now implemented in the software μ-Argus, developed under the European Union project CASC on Computational Aspects of Statistical Confidentiality. For a comprehensive approach that integrates both individual and global measures, see Polettini (2003).

In social surveys, the observed variables are frequently categorical in nature, and often comprise publicly available variables, such as sex, age, region of residence. Variables such as these that may allow identification and are accessible to the public are referred to as *key variables*. In such a framework, risk is usually defined as a function of *combinations* of values of key variables. These correspond to a contingency table built by cross-tabulating the key variables. Records presenting combinations of key variables that are unusual or rare in the population clearly have a high disclosure risk, whereas rare or even unique combinations in the sample do not necessarily correspond to high risk individuals.

Benedetti and Franconi (1998) introduced a Bayesian framework to estimate a record-level measure of re-identification risk. They noticed that $1/F_k$ is the probability of re-identification of individual i in cell k, $k = 1, \ldots, K$, when F_k individuals in the population are known to belong to this cell. In order to infer the population frequency F_k of a given combination from its sample frequency f_k, they then focused on the posterior distribution of F_k given f_k; see also Fienberg and Makov (1998). Finally, the Benedetti-Franconi risk is defined as the expected value of $1/F_k$ under this distribution. This proposal aroused a large debate that resulted in a series of papers by Di Consiglio, Franconi and Seri (2003), Polettini (2003) and Rinott (2003). In this paper we build on previous work to discuss a variety of Bayesian hierarchical models for risk estimation. For these models we derive the posterior distribution of the population frequency for each combination of values of the key variables given the observed sample frequencies. Knowledge of this distribution enables us to obtain suitable summaries that can be used to estimate the risk of disclosure; one such summary is $\mathrm{E}[1/F_k|f_k]$, but different summaries of the distribution, such as the mode or the median, can offer better performance. The methodology adopted in the paper follows a superpopulation approach similar to that used in Bethlehem, Keller and Pannekoek (1990), where a Poisson-gamma model is first proposed; Skinner and Holmes (1998) suggest instead using a Poisson-lognormal model. A different, yet related procedure is described in Carlson (2002) and Elamir and Skinner (2004).

1.1 Structure of the paper

In Section 2 we discuss in detail the approach proposed by Benedetti and Franconi (1998) and show how it is equivalent to a superpopulation model, which we call Model I. In Section 3 we present Model II, which is based on the one discussed by Bethlehem, Keller and Pannekoek (1990), and show that Model I is a limiting case of it. Model III is presented in Section 4 and is the one discussed in detail in Polettini and Stander (2004). In order to assess the risk estimates that can be obtained from each of the models, we use an artificial sample, drawn from the 1991 Italian Census data according to the sampling scheme of the Labour Force Survey, so that we know the population frequencies. This data set is discussed in Section 5. In Section 6 we refine Model III to produce Model IV. We discuss the estimated risks obtained from Model III and Model IV. Finally, in Section 7 we discuss other possible models for assessing risk in statistical disclosure limitation.

2 Model I

Let

$$\begin{aligned} \pi_k &= P(\text{a member of the population falls into cell } k), \\ p_k &= P(\text{a member of population cell } k \text{ falls into the sample}), k = 1, \ldots, K, \end{aligned}$$

where K is the number of combinations in the populations. Let the microdata file be a random sample of size n drawn from a finite population of N units.

We have already mentioned that Benedetti and Franconi (1998) estimate the risk for cell k as a posterior expectation

$$r_k = E\left(\frac{1}{F_k}\,|f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \Pr\{F_k = h\,|f_k\};$$

see Fienberg and Makov (1998), Omori (1998) and Takemura (1998). Benedetti and Franconi do not, however, formally define a hierarchical model for the population and cell frequencies. They assume that

$$F_k | f_k \sim \text{negative binomial}(f_k, p_k);$$

that is,

$$\Pr\{F_k = h | f_k = j; p_k\} = \binom{h-1}{j-1} p_k^j \left(1 - p_k\right)^{h-j}, \quad h \geq j.$$

They then estimate p_k using the sampling design weights as

$$\hat{p}_k = f_k / \hat{F}_k^D, \tag{1}$$

where $\hat{F}_k^D = \sum_{i \in \mathcal{C}_k} w_i$, in which w_i^{-1} is the probability that unit i is included in the sample and $\mathcal{C}_k$ is the set of records in the sample that belong to class k. Sometimes

the sampling weights are calibrated to match known population totals on a set of auxiliary variables, that need not be the same as the key variables; see Deville and Särndal (1992) and Di Consiglio, Franconi and Seri (2003). This can lead to problems with Benedetti and Franconi's estimate of disclosure risk; see Rinott (2003) and Di Consiglio, Franconi and Seri (2003) for a detailed discussion.

Benedetti and Franconi's assumption is equivalent to the following superpopulation model, that we shall call **Model I**:

$$\begin{aligned} \pi_k &\sim m(\pi_k) \propto 1/\pi_k,\ k = 1, \ldots, K, \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots, \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k,\ \text{independently across cells.} \end{aligned}$$

The equivalence is due to the fact that $F_k|f_k, p_k \sim$ negative binomial(f_k, p_k); see Rinott (2003) and Polettini (2003) for a more detailed discussion. Parameter estimation by an empirical Bayesian approach is not feasible for Model I as the marginal probability mass function $[f_k]$ is improper.

3 Model II

We now present another superpopulation model, which we shall call **Model II**. This model is based on the one discussed by Bethlehem, Keller and Pannekoek (1990) and takes the form

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha),\ k = 1, \ldots, K, \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots, \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k,\ \text{independently across cells,} \end{aligned}$$

in which $\alpha > 0$ is an unknown parameter. The assumption that $\pi_k \sim \text{gamma}(\alpha, K\alpha)$ means that $\text{E}[\pi_k] = 1/K$, ensuring that on average the π_ks sum to 1, and $\text{Var}[\pi_k] = 1/(K^2\alpha)$. Since K is usually very large in real applications, the very small variance means that the gamma hyperprior is very strongly concentrated on its mean, which is itself small. Hence Model II does not allow much variation across cells. It turns out that Model I can be thought of as the limit of Model II as $\alpha \to 0$ (Rinott, 2003). Hence Model I allows for more variation across cells than Model II. The same argument can be used to explain that fact that the empirical Bayesian approach for estimating the model parameters by maximizing the log-likelihood may not work well. The problem is that the marginal probability mass function $[f_k]$ tends to an improper probability mass function as $\alpha \to 0$.

4 Model III

In an attempt to allow extra variation Polettini and Stander (2004) extended Model II by modelling the p_k. Their model takes the form:

$$\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, K\alpha),\ \pi_k > 0,\ k = 1, \ldots, K, \\
F_k|\pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \\
p_k &\sim \text{beta}(a_k, b_k),\ 0 < p_k < 1, \\
f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k,\ \text{independently across cells,}
\end{aligned}$$

and will be referred to as **Model III**. Here $a_k > 0$ and $b_k > 0$ are unknown parameters. Polettini and Stander (2003) present the probability mass function of F_k given f_k and the probability mass function of f_k:

Distribution 1 *The probability mass function of F_k given f_k is*

$$[F_k|f_k] = \frac{(K\alpha)^{\alpha+f_k}\Gamma(a_k+b_k+f_k)}{\Gamma(b_k)\Gamma(\alpha+f_k)\ {}_2F_1\left(\alpha+f_k, a_k+f_k; a_k+b_k+f_k; -\frac{N}{K\alpha}\right)} \times \frac{N^{F_k-f_k}}{(N+K\alpha)^{\alpha+F_k}} \frac{\Gamma(\alpha+F_k)\Gamma(F_k-f_k+b_k)}{\Gamma(a_k+b_k+F_k)\Gamma(F_k-f_k+1)},$$

$F_k = f_k, f_k+1, \ldots,$

where ${}_2F_1\left(a, b; c; z\right)$ denotes the Hypergeometric function (Abramowitz and Stegun, 1965).

In their application Polettini and Stander (2004) only need $[F_k|f_k]$ for $f_k > 0$.

Distribution 2 *The probability mass function of f_k is*

$$[f_k] = \frac{\Gamma(b_k)}{\Gamma(\alpha)B(a_k, b_k)} \left(\frac{N}{K\alpha}\right)^{f_k} \frac{\Gamma(\alpha+f_k)\Gamma(a_k+f_k)}{\Gamma(f_k+1)\Gamma(a_k+b_k+f_k)} \times {}_2F_1\left(\alpha+f_k, a_k+f_k; a_k+b_k+f_k; -\frac{N}{K\alpha}\right),$$

$f_k = 0, 1, \ldots$.

Polettini and Stander (2004) set $a_k = a$ and $b_k = b$ so that effectively the beta distribution of p_k is not cell specific. In Section 6 we will discuss a model in which the distribution of p_k is cell specific. The empirical Bayesian approach may be problematic for Model III for the same reasons as for Model II, discussed in Section 3. In fact Polettini and Stander (2004) report problems maximizing the associated log-likelihood over (α, a, b). They do, however, arrive at estimates of these parameters that are sensible in terms of goodness of fit criteria.

5 The Data

The data that we consider are an artificial sample of $n = 53,872$ records drawn from the 1991 Italian Census data according to the sampling scheme of the Labour Force Survey, as described in Di Consiglio, Franconi and Seri (2003).

The data come from four administrative Italian regions, namely Campania, Lazio, Val d'Aosta and the Veneto. The total number of individuals in the population from these four regions is $N = 15,142,320$. Among the many variables collected in the Census, we chose the following as key variables: sex (2 categories), age (recoded in 14 classes), region of residence (the 4 regions just mentioned), position in profession (14 categories) and relationship with the head of the household (13 categories), giving $K = 2 \times 14 \times 4 \times 14 \times 13 = 20384$. Since this is an instance where the population cell frequencies F_k are known, the data allow the proposed procedure to be assessed by comparing known population quantities with their corresponding estimates.

6 Model IV

We decided to modify Model III for two reasons. First, we wanted to account for the large number of empty cells. We did this by assuming that the p_ks are drawn independently from a mixture of a beta distribution and a distribution with point mass at zero, with the weight given to the beta distribution being $\gamma \in [0, 1]$. Secondly, we wanted to make use of the $\hat{p}_k$ defined in equation (1) and used by Benedetti and Franconi (1998). If we set $a_k = a\hat{p}_k$ and $b_k = a(1 - \hat{p}_k)$ for some unknown positive parameter a to be estimated, then the beta(a_k, b_k) distribution has mean $\hat{p}_k$ and variance $\hat{p}_k(1 - \hat{p}_k)/(a + 1)$. Hence, the beta$(a_k, b_k)$ is now located around the estimated $\hat{p}_k$ and is thus cell specific. **Model IV** takes the form:

$$
\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, K\alpha),\ \pi_k > 0,\ k = 1, \ldots, K, \\
F_k|\pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots, , \\
p_k &\sim \gamma\, \text{beta}(a\hat{p}_k, a(1 - \hat{p}_k)) + (1 - \gamma)\, \delta_{\{0\}}(p_k),\ p_k \in [0, 1], \\
f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k,\ \text{independently across cells},
\end{aligned}
$$

in which the delta function $\delta_{\{0\}}$ is such that $\delta_{\{0\}}(0) = 1$ and $\delta_{\{0\}}(p_k) = 0$ for $p_k \in (0, 1]$. It is clear that when $\gamma = 1$, we recover Model III, if a_k and b_k are as just defined. The weight γ is assumed to be unknown and so has to be estimated.

It can be shown that the probability mass function $[F_k|f_k]$ remains the same as Distribution 1 with $a_k = a\hat{p}_k$ and $b_k = a(1 - \hat{p}_k)$ for $f_k > 0$. There is, however, a change to Distribution 2:

Distribution 3 *The probability mass function of f_k is now*

$$[f_k] = \begin{cases} \gamma \ {}_2F_1\left(\alpha, a\hat{p}_k; a; -\frac{N}{K\alpha}\right) + (1-\gamma) & \\ & \text{if } f_k = 0 \\ \gamma \frac{\Gamma(a(1-\hat{p}_k))}{\Gamma(\alpha)B(a\hat{p}_k, a(1-\hat{p}_k))} \left(\frac{N}{K\alpha}\right)^{f_k} \frac{\Gamma(\alpha+f_k)\Gamma(a\hat{p}_k+f_k)}{\Gamma(f_k+1)\Gamma(a+f_k)} & \\ \times \ {}_2F_1\left(\alpha+f_k, a\hat{p}_k+f_k; a+f_k; -\frac{N}{K\alpha}\right) & \\ & \text{if } f_k > 0, \end{cases}$$

It can be seen that Distribution 3 reduces to Distribution 2 when $\gamma = 1$.

6.1 Results

We adopt a fully-Bayesian type of approach, setting $\alpha = 0.1$ and $a = 80$ in Distribution 1 since $f_k > 0$. We can use the expression for $[f_k]$ given in Distribution 3 to assess goodness of fit, but we do not pursue that further here. Such an approach can guide our choice of α and a. Figure 1 shows the estimated disclosure risk obtained using Model IV and Model III, plotted against the known disclosure risk $1/F_k$. Our estimate of disclosure risk is based on the mode of the posterior distribution $[1/F_k|f_k]$, as this summary provided the best estimates. Other summaries, such as the posterior mean or median, could be considered. Model IV offers some improvement over Model III. In general, we observe the desirable feature that high risks are generally no longer underestimated. There is also a more appropriate spread in the estimated disclosure risk. Small risks tend to be overestimated, although using Model IV can reduce the extent of overestimation especially in the three large regions.

7 Discussion and Further Models

All the above models assume independence across cells. We believe that further improvements can be achieved by making some use of the structure of the contingency table. We could, for example, assume that

$$\underline{\pi} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K),$$

where $\underline{\pi} = (\pi_1, \ldots, \pi_K)$. We could then assume that

$$\begin{aligned} \underline{F}|\underline{\pi} &\sim \text{multinomial}(N; \pi_1, \ldots, \pi_K), \\ \underline{f}|\underline{F} &\sim \text{multinomial}(n; F_1/N, \ldots, F_K/N), \end{aligned}$$

for example. This approach is exactly the one suggested in Polettini and Stander (2004), except that here the assumption of equality of the parameters in the Dirichlet distribution has been relaxed. It offers the simplest way of introducing information

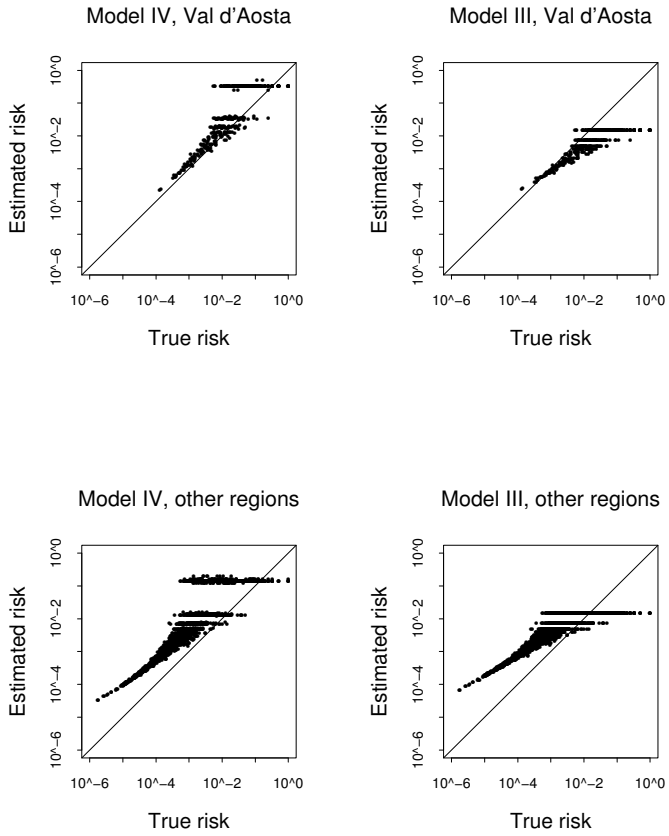

Figure 1: Scatter plots of the disclosure risks estimated using Model IV (left column) and Model III (right column) against the true risk $1/F_k$. The top row is for cells k connected with the Val d'Aosta region, while the bottom row is for three large regions Campania, Lazio and Veneto. Logarithmic scales are used for all axes.

from external archives within a Bayesian framework. When no information other than the sample is available, we suggest that the sampling design weights should be exploited by taking $\alpha_k \propto \hat{F}_k^D$. We plan to report on this aspect further. If data collected at a previous census were to be available, we could take $\alpha_k \propto F_k^{\text{previous}}$. If only marginal tables were available, we could specify a conditional independence model corresponding to these marginal tables to elicit the $(\alpha_1, \ldots, \alpha_K)$ parameters. We could also extend the above model by including the $(p_1, \ldots, p_K)$: for example,

$$\begin{aligned} \underline{p} &\sim \text{Dirichlet}(\beta_1, \ldots, \beta_K), \\ \underline{f}|\underline{F}, \underline{p} &\sim \text{multinomial}(n; p_1 F_1 / < \underline{p}, \underline{F} >, \ldots, p_K F_K / < \underline{p}, \underline{F} >), \end{aligned}$$

in which $< \underline{p}, \underline{F} >= \sum_{k=1}^{K} p_k F_k$.

The Dirichlet-multinomial approach is also proposed in a paper by Forster and Webb (2005), in which a Bayesian model averaging methodology for disclosure risk assessment is presented. It is unlikely that these models are analytically tractable, and so we may have to perform inference using Markov chain Monte Carlo methods; see Gilks, Richardson and Spiegelhalter (1996), for example. We may implement these using WinBUGS; see, for example,
`http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml` and Congdon (2001, 2003, 2005).

References

Abramowitz, M. and Stegun, I.A. (1965). *Handbook of Mathematical Functions*, Dover.

Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of New Techniques and Technologies for Statistics*, volume 1, pages 225–232, Sorrento, June 1998.

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.

Congdon, P. (2001). *Bayesian Statistical Modelling*, Wiley.

Congdon, P. (2003). *Applied Bayesian Modelling*, Wiley.

Congdon, P. (2005). *ABayesian Models for Categorical Data*, Wiley.

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. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 2003.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Busines and Economic Statistics*, **7**, 207–217.

Elamir, E. A. H. and Skinner, C. J. (2004). Modelling the re-identification risk per record in microdata. Technical report, Southampton Statistical Sciences Research Institute, University of Southampton, UK.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385–397.

Forster, J. J. and Webb, E. L. (2005). Bayesian model averaging for disclosure risk assessment. Working paper available from `http://www.maths.soton.ac.uk/staff/JJForster/paper.html`.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (Eds.) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall.

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

Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, 59–76. Eurostat, Luxembourg.

Polettini, S. (2003). Some remarks on the individual risk methodology. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 2003.

Polettini, S. and Stander, J. (2004). A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In Domingo-Ferrer, J. and Torra, V. (Eds.) *Privacy in Statistical Databases*, Berlin: Springer-Verlag, 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, 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.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 361–372.

Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, 45–58. Eurostat, Luxembourg.

Willenborg, L. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Springer.