Quaderni di Statistica
Vol. 7, 2005

Bayesian models for risk estimation in statistical disclosure limitation

Silvia Polettini
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
E-mail: spolettini@unina.it
Julian Stander
School of Mathematics and Statistics, University of Plymouth
E-mail: jstander@plymouth.ac.uk

Summary: When microdata files for research are released, it is possible that external users may attempt to breach confidentiality. For this reason most Statistical Agencies apply some form of disclosure risk assessment and data protection. Risk assessment first requires a measure of disclosure risk to be defined. The definition of disclosure risk that we adopt is based on re-identification. Therefore our risk measure is specific to cells of the contingency table built by cross tabulating the variables that allow identification. In this paper we discuss two Bayesian hierarchical models for disclosure risk estimation. Model I is an extension due to Polettini and Stander (2004) of a model discussed by Bethlehem, Keller and Pannekoek (1990). Model II is an extension of Model I that takes account of the large number of empty cells and that makes use of the available estimate of sample cell probabilities based on sampling design weights. For each model we present in detail the distributions that are necessary for risk estimation. An artificial sample of the Italian 1991 Census data allows us to assess the relative performance of each model.

Keywords: Bayesian hierarchical models, Confidentiality, Disclosure risk, Survey data.

1. Introduction

When microdata files for research are released, it is possible that external users may attempt to breach confidentiality. For this reason most

Statistical Agencies 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. 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* (see Willenborg and de Waal, 2001).

Even when attention is focused 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 Elliott (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 Franconi and Polettini (2004).

In social surveys, the observed variables are frequently categorical in nature, and often comprise publicly available variables, such as sex, age, and 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 combinations 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 (for a general Bayesian formulation of re-identification, see Fienberg and Makov, 1998). 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. Finally, they define what we refer to as the Benedetti-Franconi risk 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 analyse a model proposed in Polettini and Stander (2004), that we refer to as Model I. This model is based on the one discussed by Bethlehem, Keller and Pannekoek (1990). We then introduce an extension to Model I, that we refer to as Model II.

For each model we present all the relevant computations to derive the posterior distribution[1] of the population frequency for each combination of values of the key variables given the observed sample frequencies, $[F_k|f_1, \ldots, f_K]$. Knowledge of this distribution enables us to obtain suitable summaries that can be used to estimate the risk of disclosure; one such summary is $\mathrm{E}\,(1/F_k|f_1, \ldots, f_K)$, but different summaries, 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).

The paper is organised as follows: we begin by introducing some notation in Section 2; in Section 3 we discuss the data set that we use to assess the risk estimates that can be obtained under our approach. The data consist of an artificial sample, drawn from the 1991 Italian Census

[1]Here and in the sequel we use the compact notation $[X]$ to denote the probability mass or the probability density function of a random variable X.

data according to the sampling scheme of the Labour Force Survey, so that we know the population frequencies. In Section 4 we present Model I. In Section 5 we give all the relevant computations to obtain the posterior distribution $[F_k|f_1,\ldots,f_K]$ under Model I, whereas in Section 6 we derive the marginal law of f_k and the associated log-likelihood under the same model. In Section 7 we present Model II, which is a refinement of Model I. In this section we also derive the posterior distribution of F_k given $f_1,\ldots,f_K$, the marginal law of f_k and the associated log-likelihood under Model II. In Section 8 we discuss our approach to estimating the risk of disclosure, and in Section 9 we present the estimated risks obtained from Model I and Model II. Finally, Section 10 contains some concluding remarks and some suggestions about further models that could be used for disclosure risk estimation.

2. Some notation

Let the microdata file be a random sample of size n drawn from a finite population of N units, where N is assumed known. For a generic unit i in the population, we denote as w_i the sampling design weight, that is the reverse of the probability that i is included in the sample. We shall consider the contingency table obtained by cross-tabulating the population and the sample data according to a predefined set of key variables. This operation defines a total of K cells. We shall denote the set of records in the sample that belong to the k-th cell as $\mathcal{C}_k$.

Let

$$\pi_k = P(\text{a member of the population falls into cell } k) \qquad (1)$$

and

$$p_k = P(\text{a member of population cell } k \text{ falls into the sample}) \ , \qquad (2)$$

$k = 1,\ldots,K$.

As we want to infer the population frequencies from the sample frequencies, a distribution of interest is $[F_k|f_1,\ldots,f_k]$. This allows us to

compute the risk of disclosure for cell k

$$r_k = E\left(\frac{1}{F_k}\middle| f_1, \ldots, f_k\right). \tag{3}$$

3. 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 complex sampling scheme of the Labour Force Survey, as described in Di Consiglio, Franconi and Seri (2003). This is a widely used, unequal probability, sampling scheme.

The data come from four administrative Italian regions, namely Campania, Lazio, Val d'Aosta and 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 (14 categories), region of residence (the 4 regions just mentioned), position in profession (14 categories) and relationship with the head of the household (13 categories), giving $K = 2 \times 14 \times 4 \times 14 \times 13 = 20{,}384$. 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.

4. Defining the model

Our approach to risk estimation is based on a Bayesian hierarchical model. In order to estimate (3), we need to define a model that allows us to derive $[F_k | f_1, \ldots, f_K]$.

We assume that the π_ks are drawn independently from a gamma(α, λ) distribution:

$$[\pi_k] = \frac{\lambda^\alpha}{\Gamma(\alpha)} \pi_k^{\alpha-1} e^{-\lambda \pi_k},$$

in which $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \mathrm{d}x$. To ensure

that $E\left(\sum_{k=1}^{K} F_k\right) = N$, we impose the constraint that $E\left(\sum_{k=1}^{K} \pi_k\right) = 1$, from which it follows that $\lambda = K\alpha$. With this constraint we obtain that $E(\pi_k) = 1/K$ and $\text{Var}[\pi_k] = 1/(\alpha K^2)$.

We assume that given π_k, the F_ks are drawn independently from a Poisson$(N\pi_k)$ distribution. Therefore the probability mass function of $F_k \mid \pi_k$ is:

$$[F_k|\pi_k] = \frac{e^{-N\pi_k}\left(N\pi_k\right)^{F_k}}{F_k!}.$$

Next, we assume that the p_ks are drawn independently from a beta(a_k, b_k) distribution:

$$[p_k] = \frac{p_k^{a_k-1}\left(1-p_k\right)^{b_k-1}}{B(a_k, b_k)},$$

in which $B(a_k, b_k)$ is the beta function

$$B(a_k, b_k) = \int_0^1 x^{a_k-1}(1-x)^{b_k-1}\mathrm{d}x = \frac{\Gamma(a_k)\Gamma(b_k)}{\Gamma(a_k+b_k)}.$$

Finally, we assume that conditionally on the other random variables in the model, the observed sample cell frequencies are drawn independently from a binomial(F_k, p_k) distribution:

$$[f_k|F_k, \pi_k, p_k] = \frac{F_k!}{f_k!(F_k-f_k)!}p_k^{f_k}(1-p_k)^{F_k-f_k},$$

which does not depend on π_k.

Overall, this model, that we refer to as **Model I**, takes the following form:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha),\ \pi_k > 0,\ k = 1, \ldots, K \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots \\ p_k &\sim \text{beta}(a_k, b_k),\ 0 < p_k < 1 \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k, \end{aligned} \tag{4}$$

independently across cells.

5. Computation of $[F_k|f_1, \ldots f_K]$

The derivation of the laws $[F_k|f_1, \ldots, f_K]$ is performed in several stages. First we note that the independence assumption allows us to write $[F_k|f_1, \ldots, f_K] = [F_k|f_k]$.

For simplicity of notation we now drop the subscript k; the first step in obtaining the law $[F|f]$ is to use the following integral representation:

$$[F|f] = \int_0^\infty \int_0^1 [F|f, \pi, p][\pi|f, p][p|f]\mathrm{d}p\,\mathrm{d}\pi, \tag{5}$$

which holds since the integrand simplifies to $[F, \pi, p|f]$. In the next sections we derive explicit formulations for the elements that appear in formula (5). We shall make considerable use of the hypergeometric function (see Abramowitz and Stegun, 1965) which in its integral representation is defined as:

$${}_2F_1(A, B; C; z) = \frac{\Gamma(C)}{\Gamma(B)\Gamma(C-B)} \int_0^1 t^{B-1}(1-t)^{C-B-1}(1-tz)^{-A}\mathrm{d}t, \tag{6}$$

for $\Re(C) > \Re(B) > 0$.

5.1. Obtaining $[\pi|f, p]$

We begin by using Bayes' theorem to write

$$\begin{aligned} [\pi|f, p] &\propto [f|\pi, p][\pi|p] \\ &= [f|\pi, p][\pi] \text{ by the prior independence of } \pi \text{ and } p. \end{aligned}$$

So we need to obtain $[f|\pi, p]$.

5.2. *Obtaining* $[f|\pi,p]$

To obtain $[f|\pi,p]$ we proceed as follows:

$$\begin{aligned}[f|\pi,p] &= \sum_{F=0}^{\infty}[f|F,\pi,p][F|\pi,p] \\ &= \sum_{F=f}^{\infty}\frac{F!}{f!(F-f)!}p^f(1-p)^{F-f}e^{-N\pi}\frac{(N\pi)^F}{F!}\end{aligned}$$

given the constraint $f \leq F$

$$\begin{aligned}&= \frac{e^{-N\pi}p^f}{f!}\sum_{F=f}^{\infty}\frac{(1-p)^{F-f}(N\pi)^F}{(F-f)!} \\ &= \frac{e^{-N\pi}p^f}{f!}(N\pi)^f e^{(1-p)N\pi} \\ &= e^{-N\pi p}\frac{(N\pi p)^f}{f!},\ f=0,1,\ldots,\end{aligned}$$

by changing the variable in the summation.

Hence $f|\pi,p \sim \text{Poisson}(N\pi p)$. From $[f|\pi,p]$ it is straightforward to compute $[f|p]$, a result that we shall need soon.

5.3. *Obtaining* $[f|p]$

To obtain $[f|p]$, we write

$$[f|p] = \int_0^{\infty}[f|\pi,p][\pi]\mathrm{d}\pi;$$

using the results of the previous paragraph we have

$$\begin{aligned}
[f|p] &= \int_0^\infty e^{-N\pi p}\frac{(N\pi p)^f}{f!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\pi^{\alpha-1}e^{-\lambda\pi}\mathrm{d}\pi \\
&= \frac{(Np)^f}{f!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty \pi^{\alpha+f-1}e^{-(Np+\lambda)\pi}\mathrm{d}\pi \\
&= \frac{(Np)^f}{f!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+f)}{(Np+\lambda)^{\alpha+f}} \\
&= \frac{\Gamma(\alpha+f)}{\Gamma(\alpha)\Gamma(f+1)}\left(\frac{\lambda}{Np+\lambda}\right)^\alpha\left(\frac{Np}{Np+\lambda}\right)^f, \qquad (7)
\end{aligned}$$

$f = 0, 1, \ldots,$ where we write $f! = \Gamma(f+1)$.

In fact, if $\alpha \geq 1$ is an integer, this is a negative binomial distribution, representing the probability that f `tail`s are thrown before the α^{th} `head`, where $P(\texttt{head}) = \lambda/(Np+\lambda) = K\alpha/(Np+K\alpha)$.

5.4. Obtaining $[F]$

The marginal of F takes a similar form:

$$[F] = \int_0^\infty [F|\pi][\pi]\mathrm{d}\pi;$$

substituting the expressions for $[F|\pi]$ and $[\pi]$ defined in (4) we obtain:

$$\begin{aligned}
[F] &= \int_0^\infty [F|\pi][\pi]\mathrm{d}\pi \\
&= \int_0^\infty e^{-N\pi}\frac{(N\pi)^F}{F!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\pi^{\alpha-1}e^{-\lambda\pi}\mathrm{d}\pi \\
&= \frac{N^F}{F!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty \pi^{\alpha+F-1}e^{-(N+\lambda)\pi}\mathrm{d}\pi \\
&= \frac{N^F}{F!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+F)}{(N+\lambda)^{\alpha+F}} \\
&= \frac{\Gamma(\alpha+F)}{\Gamma(\alpha)\Gamma(F+1)}\left(\frac{\lambda}{N+\lambda}\right)^\alpha\left(\frac{N}{N+\lambda}\right)^F, \; F = 0, 1, \ldots.
\end{aligned}$$

5.5. Returning to $[\pi|f,p]$

Now that we have $[f|\pi,p]$, we can easily obtain $[\pi|f,p]$:

$$\begin{aligned}[\pi|f,p] &\propto [f|\pi,p][\pi] \\ &= e^{-N\pi p}\frac{(N\pi p)^f}{f!}\frac{\lambda^\alpha}{\Gamma(\alpha)}\pi^{\alpha-1}e^{-\lambda\pi} \\ &\propto \pi^{\alpha+f-1}e^{-(Np+\lambda)\pi} \text{ after some simplification.}\end{aligned}$$

From the previous equation we can immediately recognize that $\pi|f,p \sim$ gamma$(\alpha+f, Np+\lambda)$, from which we obtain the full form of the law:

$$[\pi|f,p] = \frac{(Np+\lambda)^{\alpha+f}}{\Gamma(\alpha+f)}\pi^{\alpha+f-1}e^{-(Np+\lambda)\pi},\ \pi>0.$$

5.6. Obtaining $[p|f]$

The next ingredient that we require for the calculation of $[F|f]$ is $[p|f]$. It turns out that this requires the hypergeometric function. We begin simply:

$$\begin{aligned}[p|f] &\propto [f|p][p] \\ &= \frac{\Gamma(\alpha+f)}{\Gamma(\alpha)\Gamma(f+1)}\left(\frac{\lambda}{Np+\lambda}\right)^\alpha\left(\frac{Np}{Np+\lambda}\right)^f\frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} \\ &\propto \frac{p^{a+f-1}(1-p)^{b-1}}{(Np+\lambda)^{\alpha+f}},\ 0<p<1.\end{aligned}$$

We need the multiplicative normalizing constant c:

$$c = \left\{\int_0^1 \frac{p^{a+f-1}(1-p)^{b-1}}{(Np+\lambda)^{\alpha+f}}\mathrm{d}p\right\}^{-1}. \tag{8}$$

If we make the following substitutions into the hypergeometric function (6):

$$\begin{aligned} A &= \alpha + f \in \Re \\ B &= a + f \in \Re \text{ and } B > 0 \\ C &= a + b + f \in \Re \text{ and } C > B \\ z &= -\frac{N}{\lambda} \in \Re \end{aligned}$$

we obtain after a little simplification

$$\begin{aligned} {}_2F_1 &\left(\alpha + f, a + f; a + b + f; -\frac{N}{\lambda}\right) \\ &= \lambda^{\alpha+f} \frac{\Gamma(a+b+f)}{\Gamma(a+f)\Gamma(b)} \int_0^1 \frac{t^{a+f-1}(1-t)^{b-1}}{(Nt+\lambda)^{\alpha+f}} \mathrm{d}t, \end{aligned}$$

so that

$$c = \lambda^{\alpha+f} \frac{\Gamma(a+b+f)}{\Gamma(a+f)\Gamma(b)} \frac{1}{{}_2F_1\left(\alpha + f, a + f; a + b + f; -\frac{N}{\lambda}\right)}.$$

If we now write

$$\mathcal{H}(f, \alpha, a, b) = {}_2F_1\left(\alpha + f, a + f; a + b + f; -\frac{N}{\lambda}\right) \tag{9}$$

for simplicity of notation, we obtain

$$c = \lambda^{\alpha+f} \frac{\Gamma(a+b+f)}{\Gamma(a+f)\Gamma(b)} \frac{1}{\mathcal{H}(f, \alpha, a, b)} \tag{10}$$

and so

$$[p|f] = \frac{\lambda^{\alpha+f}\Gamma(a+b+f)}{\Gamma(a+f)\Gamma(b)\mathcal{H}(f, \alpha, a, b)} \frac{p^{a+f-1}(1-p)^{b-1}}{(Np+\lambda)^{\alpha+f}}, \qquad 0 < p < 1.$$

The final ingredient in the calculation of $[F|f]$ is $[F|f, \pi, p]$.

5.7. Obtaining $[F|f,\pi,p]$

By Bayes' theorem we have

$$[F|f,\pi,p] = [f|F,\pi,p]\frac{[F|\pi,p]}{[f|\pi,p]} = [f|F,\pi,p]\frac{[F|\pi]}{[f|\pi,p]},$$

since given π, F is independent of p.

Using the expression obtained above for $[f|\pi,p]$ we finally get:

$$\begin{aligned}[F|f,\pi,p] &= \frac{F!}{f!(F-f)!}p^f(1-p)^{F-f}\frac{e^{-N\pi}(N\pi)^F}{F!}\frac{f!}{e^{-N\pi p}(N\pi p)^f} \\ &= \frac{(1-p)^{F-f}(N\pi)^{F-f}e^{-N\pi(1-p)}}{(F-f)!},\end{aligned}$$

$F = f, f+1, \ldots$.

We now have everything that we need to find $[F|f]$.

5.8. Putting everything together to obtain $[F|f]$

In Section 5 we wrote

$$[F|f] = \int_0^\infty \int_0^1 [F|f,\pi,p][\pi|f,p][p|f]\mathrm{d}p\,\mathrm{d}\pi.$$

So substituting the above results for $[F|f,\pi,p]$, $[\pi|f,p]$ and $[p|f]$ we obtain

$$\begin{aligned}
[F|f] &= \int_0^\infty \int_0^1 \frac{(1-p)^{F-f}(N\pi)^{F-f}e^{-N\pi(1-p)}}{(F-f)!}\frac{(Np+\lambda)^{\alpha+f}}{\Gamma(\alpha+f)}\times \\
&\qquad \pi^{\alpha+f-1}e^{-(Np+\lambda)\pi}c\frac{p^{a+f-1}(1-p)^{b-1}}{(Np+\lambda)^{\alpha+f}}\mathrm{d}p\,\mathrm{d}\pi \\
&= \frac{N^{F-f}c}{(F-f)!\Gamma(\alpha+f)}\times \\
&\qquad \int_0^\infty \int_0^1 p^{a+f-1}(1-p)^{F-f+b-1}\pi^{\alpha+F+-1}e^{-\pi(N+\lambda)}\mathrm{d}p\,\mathrm{d}\pi \\
&= \frac{N^{F-f}c}{(F-f)!\Gamma(\alpha+f)}\int_0^\infty \pi^{\alpha+F-1}e^{-\pi(N+\lambda)}\mathrm{d}\pi\times \\
&\qquad \int_0^1 p^{a+f-1}(1-p)^{F-f+b-1}\mathrm{d}p \\
&= \frac{N^{F-f}c}{(F-f)!\Gamma(\alpha+f)}\frac{\Gamma(\alpha+F)}{(N+\lambda)^{\alpha+F}}B(a+f,F-f+b).
\end{aligned}$$

Substituting in for c from (10) we get

$$\begin{aligned}
[F|f] = \frac{N^{F-f}}{(F-f)!\Gamma(\alpha+f)}\frac{\Gamma(\alpha+F)}{(N+\lambda)^{\alpha+F}}\times \\
B(a+f,F-f+b)\lambda^{\alpha+f}\frac{\Gamma(a+b+f)}{\Gamma(a+f)\Gamma(b)}\frac{1}{\mathcal{H}(f,\alpha,a,b)}.
\end{aligned}$$

Writing the beta function in terms of gamma functions, rearranging the terms and reintroducing the subscripts we finally get the following result.

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

$$\begin{aligned}
[F_k|f_k] = \frac{\lambda^{\alpha+f_k}\Gamma(a_k+b_k+f_k)}{\Gamma(b_k)\Gamma(\alpha+f_k)\mathcal{H}(f_k,\alpha,a_k,b_k)}\frac{N^{F_k-f_k}}{(N+\lambda)^{\alpha+F_k}}\times \\
\frac{\Gamma(\alpha+F_k)\Gamma(F_k-f_k+b_k)}{\Gamma(a_k+b_k+F_k)\Gamma(F_k-f_k+1)}, F_k = f_k, f_k+1,\ldots.
\end{aligned}$$

6. Computation of $[f_k]$

It is now straightforward to find the marginal distribution $[f]$ using the expression for $[f|p]$:

$$\begin{aligned}
[f] &= \int_0^1 [f|p][p]\mathrm{d}p \\
&= \int_0^1 \frac{\Gamma(\alpha+f)}{\Gamma(\alpha)\Gamma(f+1)} \left(\frac{\lambda}{Np+\lambda}\right)^{\alpha} \left(\frac{Np}{Np+\lambda}\right)^{f} \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}\mathrm{d}p \\
&= \frac{\Gamma(\alpha+f)\lambda^{\alpha}N^{f}}{\Gamma(\alpha)\Gamma(f+1)B(a,b)} \int_0^1 \frac{p^{a+f-1}(1-p)^{b-1}}{(Np+\lambda)^{\alpha+f}}\mathrm{d}p \\
&= \frac{\Gamma(\alpha+f)\lambda^{\alpha}N^{f}}{\Gamma(\alpha)\Gamma(f+1)B(a,b)} \frac{1}{c},
\end{aligned}$$

by the definition of c given in (8). Using the expression for c given in (10) and after some simplification we finally have:

Distribution 2 *The probability mass function of* f_k *for* $k = 1, \ldots, K$ *is*

$$[f_k] = \frac{\Gamma(b_k)}{\Gamma(\alpha)B(a_k,b_k)} \left(\frac{N}{\lambda}\right)^{f_k} \frac{\Gamma(\alpha+f_k)\Gamma(a+f_k)}{\Gamma(f_k+1)\Gamma(a_k+b_k+f_k)} \mathcal{H}(f_k,\alpha,a_k,b_k),$$

$f_k = 0, 1, \ldots$

This marginal distribution can be used for goodness of fit purposes and to compute the likelihood.

6.1. The log-likelihood function

In order to estimate the model parameters by maximum likelihood, thus performing an empirical Bayesian (EB) analysis (see Efron and Morris, 1973), we can now consider the log-likelihood function of α, $a_1, \ldots, a_K$ and $b_1, \ldots, b_K$ given data $f_1, \ldots, f_K$. Up to an additive constant this can

be written as

$$
\begin{aligned}
L(\alpha, a_1, \ldots, a_K, b_1, \ldots, b_K) &= -K \log \Gamma(\alpha) \\
&+ \sum_{k=1}^{K} \Bigg\{ \log \Gamma(a_k + b_k) - \log \Gamma(a_k) + \log \Gamma(\alpha + f_k) \\
&\quad + \log \Gamma(a_k + f_k) - \log \Gamma(a_k + b_k + f_k) - f_k \log \alpha \\
&\quad + \log {}_2F_1 \left(\alpha + f_k, a_k + f_k; a_k + b_k + f_k; -\frac{N}{K\alpha} \right) \Bigg\} \quad .
\end{aligned}
\tag{11}
$$

Model (4) is cell-specific, because the parameters a_k and b_k are allowed to depend on the cell k. To obtain maximum likelihood estimates of the parameters, Polettini and Stander (2004) make the simplifying assumption that $a_1 = \cdots = a_K = a$ and $b_1 = \cdots = b_K = b$. This is a special case of Model I, where

$$
\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, \lambda),\ \pi_k > 0,\ k = 1, \ldots, K, \\
F_k | \pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots, \\
p_k &\sim \text{beta}(a, b),\ 0 < p_k < 1, \\
f_k | F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k,
\end{aligned}
\tag{12}
$$

independently across cells at all levels. Up to an additive constant, the log-likelihood function (11) then becomes

$$
\begin{aligned}
L(\alpha, a, b) &= -K \left\{ \log \Gamma(\alpha) - \log \Gamma(a + b) + \log \Gamma(a) \right\} \\
&+ \sum_{k=1}^{K} \Bigg\{ \log \Gamma(\alpha + f_k) + \log \Gamma(a + f_k) \\
&\quad - \log \Gamma(a + b + f_k) - f_k \log \alpha \\
&\quad + \log {}_2F_1 \left(\alpha + f_k, a + f_k; a + b + f_k; -\frac{N}{K\alpha} \right) \Bigg\} \quad .
\end{aligned}
\tag{13}
$$

7. An extension of Model I

Model (12) has the drawback that all cells having the same sample frequency f_k will have the same risk, since the posterior distribution $[F_k | f_k]$

only depends on cell k through f_k. This is not a desirable feature, as we want to be able to classify records – sample uniques, for example – into safe and unsafe.

We decided to modify the model estimated in Polettini and Stander (2004) for several reasons. First, we wanted to define the model so that the risk was cell-specific; secondly, we wanted to make use of the sampling design weights that are released with the data. For this aim we introduced the $\hat{p}_k$ used by Benedetti and Franconi (1998) and defined as

$$\hat{p}_k = \frac{f_k}{\sum_{i \in \mathcal{C}_k} w_i}, \tag{14}$$

where, as mentioned in Section 2, w_i is the sampling design weight attached to record i. Finally, we wanted to account for the large number of empty cells: in practical applications there is indeed a large number of empty cells, many of which derive from structural zeroes in the population contingency table. These are not accounted for by Model I. We took account of these aspects 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 weight $\gamma \in [0, 1]$.

Model II takes the form:

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

independently across cells, in which the delta function $\delta_{\{0\}}$ is such that $\delta_{\{0\}}(0) = 1$ and $\delta_{\{0\}}(p_k) = 0$ for $p_k \in (0, 1]$.

We have therefore imposed on the distribution of p_k the constraint that it has mean $\hat{p}_k$. That was achieved by setting $a_k = a\hat{p}_k$ and $b_k = a(1-\hat{p}_k)$ for some unknown positive parameter a. Under this parametrisation, the beta(a_k, b_k) is now located around the estimated $\hat{p}_k$ with variance $\hat{p}_k(1-\hat{p}_k)/(a+1)$ and is thus cell specific.

The weight γ is not further specified and so has to be elicited or estimated. It is clear that when $\gamma = 1$, we recover Model I, if a_k and b_k are as just defined.

Besides returning a cell-specific risk measure, this specification also partially accounts for the presence of different sampling fractions in different cells, a characteristic that is typical of most sampling designs. In our application, for instance, in order to obtain estimates that have approximately the same precision across region in the presence of both small regions and large regions, the sampling fraction varies considerably across regions.

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 the marginal distribution of f_k reported in Distribution 2:

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

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

In fact we do not have the value of $\hat{p}_k$ for cells with $f_k = 0$; we will discuss this further in Section 7.1.

7.1. The log-likelihood function

The log-likelihood $L(\alpha, a, \gamma)$ now takes the form

$$\left\{ \sum_{k:\, f_k > 0} \log [f_k] \right\} + (K - \text{number of non-empty cells}) \log[f_k = 0];$$

in our application $K = 20,384$ and there are $2,966$ non-empty cells. As mentioned in Section 7, we do not have $\hat{p}_k$s for cells with $f_k = 0$. To overcome this, we set

$$\begin{aligned} \log[f_k = 0] \;\; = \;\; & 0.86 \log \left\{ \gamma \, {}_2F_1\left(\alpha, 0.0034a; a; -\frac{N}{K\alpha}\right) + (1-\gamma) \right\} + \\ & 0.14 \log \left\{ \gamma \, {}_2F_1\left(\alpha, 0.028a; a; -\frac{N}{K\alpha}\right) + (1-\gamma) \right\}. \end{aligned}$$

We adopted this choice because the histogram of the $\hat{p}_k$s corresponding to cells with $f_k > 0$ was clearly bimodal, with one mode centered on 0.028 corresponding to cells connected with Val d'Aosta, and the other centered on the lower 0.0034 corresponding to cells connected with the three other regions (the larger Campania, Lazio and Veneto). The first mode comprised 14% of the p_ks, while the second mode comprised the remaining 86%. The charachteristics of the sampling design reported in Section 7 imply that the $\hat{p}_k$s corresponding to cells connected with Val d'Aosta are considerably greater than that for the three large regions.

8. Our approach to estimating r_k

As mentioned, the risk of disclosure for records in cell k is defined in terms of the posterior distribution of the population cell frequency F_k given the observed data, $f_1, f_2, \ldots, f_K$.

Under Models I and II, the assumption of independence allows us to restrict attention to $[F_k|f_k]$ to compute

$$r_k = E\left(\frac{1}{F_k}\middle| f_k\right) = \sum_{h=f_k}^{\infty} \frac{1}{h} \Pr\left(F_k = h|f_k\right), \tag{15}$$

where the conditional distribution $[F_k|f_k]$ depends on the unknown parameters of the model.

Without further assumptions on the hyperparameters of the model, the risk of disclosure for cell k cannot be evaluated under model (4), as the risk depends on $2K + 1$ unknown parameters, namely α, a_k and b_k, $k = 1, \ldots, K$. Under model (12), knowledge of $[f_k]$ may allow us to estimate the unknown parameters of the model α, a and b from the data $f_1, \ldots, f_K$ via an EB approach. The risk (15) is then estimated by plugging in the estimates of the unknown parameters in the model.

As the results in Polettini and Stander (2004) illustrate, the EB approach does not work well under Model I. This is because the probability mass function $[f_k]$ tends to an improper distribution when $\alpha \to \infty$, so that the likelihood function diverges when α tends to zero.

The same problem arises under Model II. For this reason in the present

paper we adopt a fully Bayesian approach, therefore eliciting the prior distributions completely, using the marginal $[f_k]$ to assess our elicitation.

We set $\alpha = 0.92$, $a = 0.86$ and $b = 80$ in Model I and $\alpha = 0.1$, $a = 80$ and $\gamma = 0.7$ in Model II.

9. Results

We compare the performance of the two models discussed above by using the data described in Section 3. As an assessment of the procedure, we show in Figure 1 the estimated disclosure risk obtained using Model I and Model II, plotted against the known disclosure risk $1/F_k$. Model II

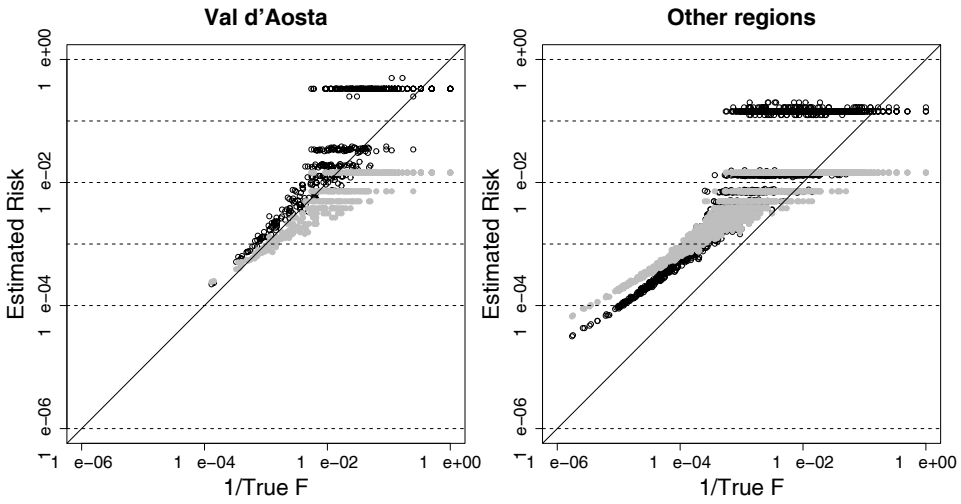

Figure 1: Scatter plots of the disclosure risks estimated using Model I (in grey) and Model II (in black) against the true risk $1/F_k$. The left panel is for the Val d'Aosta region, while the right panel is for three large regions Campania, Lazio and Veneto. Logarithmic scales are used for all axes.

offers some improvement over Model I. 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 II can reduce the

extent of overestimation especially in the three large regions. Some bias remains, especially for the large regions.

Following Forster (2005), we consider the procedure of risk assessment as a classifier, defining a cell as unsafe if its risk is greater than 0.05. Tables 1 and 2 show how the classifier performs under both models. It can be seen that Model II is a great improvement over Model I.

	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	312	109
$\hat{r}_k > 0.05$	0	0

Table 1: Performance of Model I as a classifier

	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	231	5
$\hat{r}_k > 0.05$	81	104

Table 2: Performance of Model II as a classifier

10. Conclusion and discussion

In order for Statistical Agencies to perform risk assessment and data protection, measures of disclosure risk are needed. These measures can be used to screen out unsafe records and apply protection selectively. The results reported in this paper indicate that the proposed methodology is sensible; experiments not reported in this paper also show that Model II offers some improvement over the risk estimation procedure currently implemented in the software μ-Argus (see Franconi and Polettini, 2004).

Both the models that we have studied assume independence across cells, although centering the distribution of the p_ks on $\hat{p}_k$ relaxes this assumption. Indeed the $\hat{p}_k$s depend on calibrated sampling design weights (see Deville and Särndal, 1992), so that effectively Model II takes account of the association structure of the population contingency table on which

the sampling weights are calibrated through the prior distribution on p_k. We believe that further improvements can be achieved by making full use of the structure of the contingency table. Polettini and Stander (2004) suggested a Dirichlet-multinomial-multinomial framework for this. Again, the problem of eliciting hyperparameters arises. For this model it is not possible to obtain the corresponding conditional and marginal distributions in closed form and so inference has to be performed using Markov chain Monte Carlo simulations. We have already gained some experience with this model, the results from which seem quite promising. We plan to report this in full detail in another paper.

Acknowledgements: The authors would like to thank Luisa Franconi for helpful conversations and a referee for useful comments on the paper.
The authors gratefully acknowledge the financial support of the European Union project IST-2000-25069 CASC on Computational Aspects of Statistical Confidentiality.

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, 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.

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 *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality*, Eurostat, Luxembourg, 286-298.

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

Efron B. and Morris C. (1973), Stein's estimation rule and its competitors

-an empirical Bayes approach, *Journal of the American Statistical Association*, 68, 117-130.

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. (2005), Bayesian methods for disclosure risk assessment, in *Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality*, Geneva, Switzerland, 9-11 November 2005.
http://www.unece.org/stats/documents/2005.11.confidentiality.htm

Franconi L. and Polettini S. (2004), Individual risk estimation in μ-Argus: a review, in Domingo-Ferrer, J. and Torra, V. (Eds.) *Privacy in Statistical Databases*, Berlin: Springer-Verlag, 262-272.

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

Polettini S. (2003), Some remarks on the individual risk methodology, in *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality*, Eurostat, Luxembourg, 299-311.

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 *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality*, Eurostat, Luxembourg, 275-285.

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.

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