

A Bayesian approach to estimating the risk of disclosure for survey data

*Stima del rischio di violazione della riservatezza in indagini
socio-demogra che: un approccio bayesiano.*

Silvia Polettini¹ and Julian Stander²

¹ Dipartimento di Scienze Statistiche, Università di Napoli Federico II
e-mail: spolettini@unina.it

² School of Mathematics and Statistics, University of Plymouth, UK

Riassunto: Nella fase di rilascio dei dati, il vincolo alla tutela della riservatezza dei rispondenti impone la valutazione del rischio di violazione della riservatezza e, ove necessario, la protezione dei dati. Riferendoci esclusivamente al rilascio di dati elementari, utilizziamo il concetto di reidentificazione per definire una misura di rischio a livello individuale. Essa risulta funzione delle frequenze delle celle della tabella di contingenza, usualmente non nota, che classifica le unità della popolazione secondo le variabili ritenute utili alla reidentificazione. Per stimare il rischio utilizziamo un approccio bayesiano, introducendo alcuni modelli di superpopolazione che vengono confrontati su dati simulati.

Keywords: Bayesian hierarchical models, confidentiality, disclosure risk, MCMC.

1. Introduction

When microdata files for research are released, malicious users may attempt to breach confidentiality. Disclosure risk assessment allows us to judge whether a particular data release is safe; when this is considered unsafe, protection measures must be put into practice. 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. We focus on record-level risk measures, as in Benedetti and Franconi (1998), Skinner and Holmes (1989) and Elamir and Skinner (2004), for example. We define the risk of record i as the probability of it being correctly matched to a population record.

In social surveys, often the observed variables are categorical and comprise publicly available variables. If these allow identification, they are referred to as *key variables*. Given our definition of disclosure, the risk is specific to cells (or *combinations*) of the contingency table built by cross-tabulating the key variables. Records presenting combinations of key variables that are 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. In a Bayesian framework, we define the risk as the predictive probability of a correct match given the sample cell frequencies $\underline{f} = (f_1, \dots, f_K)$. When F_k individuals in the population are known to belong to this cell, the probability of re-identification of individual i in cell k is $1/F_k$. We measure the risk as the expected value of $1/F_k$ given the sample contingency table, $r_k = E(1/F_k | \underline{f})$. This definition is similar to that used by Benedetti and Franconi (1998); to infer F_k , they refer

to the posterior distribution of F_k given f_k , defining the risk as $r_k = E(1/F_k|f_k)$. See Rinott (2003) and references therein for discussion of this approach.

2. A Poisson-gamma Model

We present a Bayesian hierarchical model that extends work by Bethlehem, Keller and Pannekoek (1990) and Rinott (2003); see also Polettini and Stander (2004). The basic component is a Poisson-gamma model for each population cell frequency F_k , followed by a conditionally binomial model for the sample cell frequency f_k s given the rest. We noticed that the constraints imposed by the original formulation do not allow enough variation across cells, and decided to model the binomial probabilities. In such a framework we also wanted to account for the large number of empty cells in our contingency tables. Following Skinner and Holmes (1993), we assumed that the p_k s are drawn independently from a mixture of a beta distribution and a point mass at zero. Secondly, we always deal with complex, unequal probability, sampling schemes that we try to approximate by assuming $f_k|F_k, p_k \sim \text{binomial}(F_k, p_k)$. These models could be informed by using the known sampling weights w_i . One possible choice is $\hat{p}_k = f_k/\hat{F}_k^D$ (see Benedetti and Franconi, 1998), where $\hat{F}_k^D = \sum_{i \in \mathcal{C}_k} w_i$, and $\mathcal{C}_k$ is the set of records in the sample that belong to cell k . A reparametrisation of the $\text{beta}(a_k, b_k)$ distribution allows us to locate it around $\hat{p}_k$, so as to make the model cell specific. We set $a_k = v\hat{p}_k$ and $b_k = v(1 - \hat{p}_k)$ for some positive parameter v . **Model I** takes the form:

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

independently across cells.

Under this model the conditional distribution of $F_k|f_1, \dots, f_k$ only depends on the corresponding sample cell frequency f_k . For $f_k > 0$ and a_k and b_k as just defined, the probability mass function of F_k given f_k is

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

${}_2F_1(a, b; c; z)$ denoting the Hypergeometric function, or Gauss series.

3. A Dirichlet-multinomial model

Model I assumes independence across cells and does not use the structure of the contingency table, except possibly when an empirical Bayesian approach is used to specify the prior for the p_k s. Setting $\underline{\pi} = (\pi_1, \dots, \pi_K)$, an alternative model could be

$$\begin{aligned} \underline{\pi} &\sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K), \\ \underline{F}|\underline{\pi} &\sim \text{multinomial}(N; \pi_1, \dots, \pi_K), \\ \underline{f}|\underline{F} &\sim \text{multinomial}(n; F_1/N, \dots, F_K/N). \end{aligned}$$

This approach is the same as in Polettini and Stander (2004), except that the symmetry of the Dirichlet distribution has been relaxed. Forster and Webb (2005) also propose a Dirichlet-multinomial approach, in a Bayesian model averaging framework. We refer to this Dirichlet-multinomial-multinomial model as **Model II**. Information from external archives can be introduced through the prior on $\underline{\pi}$. When no information other than the sample is available, we suggest that the sampling design weights could be exploited by taking $\alpha_k \propto \hat{F}_k^D$. 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 log-linear model corresponding to these marginal tables to elicit the $(\alpha_1, \dots, \alpha_K)$ parameters. We used WinBUGS to perform inference using Markov chain Monte Carlo methods. We have also written our own routines in R for confirmation.

4. Application and discussion

We consider an artificial sample drawn from the 1991 Italian Census data according to the complex sampling scheme of the Labour Force Survey (see Di Consiglio et al., 2003 for details). Since the population cell frequencies F_k are known here, the data allow the proposed procedures to be assessed by comparing known population quantities with their corresponding estimates.

The data come from four administrative Italian regions, namely Campania, Lazio, Val d'Aosta and Veneto. We chose the following key variables: sex, age in classes, region of residence, position in profession and relationship with the head of the household. We present results for Val d'Aosta only. Compared to the other regions, Val d'Aosta has a large number of potentially unsafe cells; only a small portion of these, however, corresponds to small or unique cells in the population. This feature is due to the sampling design that requires higher sampling rates for Val d'Aosta to achieve estimates with the same precision across regions.

Figure 1: Risks estimated under Model II vs. true risks. From left to right: $\alpha_k = F_k$, the fitted values from a saturated log-linear model; $\alpha_k = \hat{F}_k$, the fitted values from log-linear model 1; $\alpha_k = \hat{F}_k$, the fitted values from log-linear model 2; $\alpha_k = \hat{F}_k^D$, the design-based estimates. Estimates from Model I have been superimposed in grey.

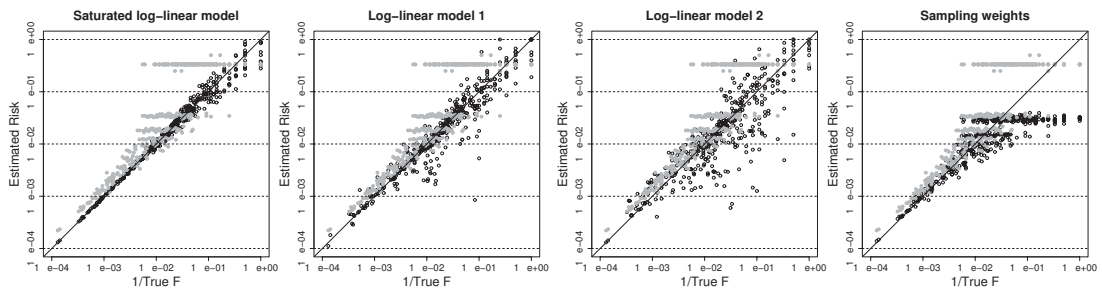


Figure 1 shows the estimates obtained from Model II; risk estimates from Model I are superimposed in grey. Overall, Model II seems to perform better than Model I, especially for small and hence potentially high risk cells. The first panel of Figure 1 was produced using $\alpha_k = F_k$. These are in fact the fitted-values from a saturated log-linear model.

Estimation is clearly very good, even for high values of disclosure risk. This result is overly optimistic, as the F_k s would not be known in practice. The second and third panel were produced using $\alpha_k = \hat{F}_k$, the fitted values from the following log-linear models:

- log-linear model 1: $F \sim \text{sex} + (\text{rel} + \text{age} + \text{posprof})^3$;
- log-linear model 2: $F \sim \text{rel} + (\text{sex} + \text{age} + \text{posprof})^3$.

Not unexpectedly, in both cases the performance is less good than before. Finally, the fourth panel of Figure 1 was produced using $\alpha_k = \hat{F}_k^D$, the estimates based on the sampling design weights. Performance is quite poor here, high risks being considerably under-estimated.

Our assessment of the performance of Model II is optimistic because we use the population frequencies that we are trying to estimate to inform the Dirichlet prior. The idea is, however, to use data from a previous census, if it were accessible. We believe that such data will soon be made available to us. Finding ways of using all the relevant archive information is a challenging problem.

Often, the assumption of a Dirichlet prior in the analysis of contingency tables is criticised as being too restrictive; see O’Hagan and Forster (2004). If the prior variance is too low for example, robustness to prior specification may be lacking. In fact, our application shows that the specification of the prior has a large impact on the risk estimates. Accordingly, we plan to examine ways of relaxing the Dirichlet assumption. Finally, we need to improve the performance of our MCMC samplers.

References

- Benedetti R., 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., Pannekoek J. (1990) Disclosure control of microdata, *Journal of the American Statistical Association*, 85, 38–45.
- Di Consiglio L., Franconi L., 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.
- Elamir E. A. H., Skinner C. J. (2004) Modelling the re-identification risk per record in microdata, Technical report, Southampton Statistical Sciences Research Institute, University of Southampton, UK.
- Forster J. J., Webb E. L. (2005) Bayesian model averaging for disclosure risk assessment, Working paper, URL <http://www.maths.soton.ac.uk/staff/JJForster/paper.html>.
- O’Hagan A., Forster J. J. (2004) Bayesian Inference, 2nd edition, volume 2B of *Kendall’s Advanced Theory of Statistics*, Arnold, London.
- Polettini S., Stander J. (2004) A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation, in: Domingo-Ferrer J., Torra V. (Eds.) *Privacy in Statistical Databases*, page 247–261, Springer-Verlag, Berlin.
- 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., Holmes D. J. (1998) Estimating the re-identification risk per record in microdata, *Journal of Official Statistics*, 14, 361–372.