WP. 12
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

BAYESIAN METHODS FOR DISCLOSURE RISK ASSESSMENT

Invited Paper

Submitted by the University of Southampton, United Kingdom[1]

[1] Prepared by Jonathan J. Forster (j.j.forster@soton.ac.uk).

Bayesian methods for disclosure risk assessment

Jonathan J Forster*

* Southampton Statistical Sciences Research Institute, School of Mathematics, University of Southampton, UK. (J.J.Forster@soton.ac.uk)

Abstract. Measures of the risk of individual identification in the release of categorical microdata are commonly based on the probability of an intruder correctly matching an individual in the population to a record in the released data. In this paper, we discuss how such probabilities can be interpreted and focus on Bayesian predictive probabilities as risk measures. By utilising a Bayesian approach to estimation under model uncertainty, known as model-averaging, we can provide more realistic estimates of disclosure risk for individual records than are provided by methods which ignore the multivariate structure of the data set. The method is illustrated with an example.

1 Introduction

Suppose that an agency releases categorical data on a sample of individuals from a population. Then, the sample data can be expressed as a a multiway contingency table. Identification risk occurs when there are sample cell counts of 1 (uniques) in the marginal table representing the cross-classification of individuals by a subset of *key* variables (those variables whose values in the population are available to a potential intruder from a source external to the released data under consideration). If the intruder can determine, with confidence, that a sample unique in the contingency table of key variables is also unique in the population, then this individual can be identified and the data release allows disclosure of the values of the remaining (non-key) variables for this individual. In this paper we focus on the disclosure risk associated with the release of individual records as part of a larger database.

It is common to quantify individual record disclosure risk as the probability of an individual being identified. Let $f_1, \dots, f_K$ denote the sample cell counts in the contingency table of key variables and $F_1, \dots, F_K$ the corresponding population cell counts and let n and N represent the sample and population totals respectively. Then, *given the intruder had knowledge of the population cell counts* $\boldsymbol{F} = (F_1, \dots, F_K)$, they could match any individual in the population whose record belonged in cell j, with a record chosen from the sample, and (in the absence of any other information) would be able to evaluate the probability of a correct match as

$1/F_j$. We denote this as the conditional probability

$$P(E_j|\boldsymbol{F}) = \frac{1}{F_j} \tag{1}$$

where E_j is used to denote the event that an individual whose record is in cell j is correctly matched. This measure was proposed by Benedetti and Franconi (1998). In practice, the intruder only has knowledge of the sample cell counts $\boldsymbol{f} = (f_1, \ldots, f_K)$ and cannot calculate (1). One alternative is to use the sample data to estimate the F_j in (1). Alternatively,it might be considered that the relevant disclosure risk measure is $P(E_j|\boldsymbol{f})$, the predictive probability of a correct match given only the sample data. The easiest interpretation of this quantity is within a Bayesian statistical framework, as follows.

Bayesian inference uses probability to quantify uncertainty. Hence, the uncertainty about any unknowns prior to obtaining sample data is encapsulated in a prior probability distribution. On observing data, this is then updated to a posterior distribution using Bayes theorem. In the present context, the unknowns are the population counts $\boldsymbol{F}$, the sample data are $\boldsymbol{f}$ and we obtain the posterior distribution as

$$P(\boldsymbol{F}|\boldsymbol{f}) \propto P(\boldsymbol{F})P(\boldsymbol{f}|\boldsymbol{F}) \tag{2}$$

where $P(\boldsymbol{F})$ is the prior distribution for $\boldsymbol{F}$ and $P(\boldsymbol{f}|\boldsymbol{F})$ represents the sampling distribution of the observed table. Having observed the sampled records it is only the unsampled records about which uncertainty remains, hence (2) can be replaced by

$$P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f}) \propto P(\boldsymbol{F}-\boldsymbol{f})P(\boldsymbol{f}|\boldsymbol{F}-\boldsymbol{f}) \tag{3}$$

Now, we simply note that our required disclosure risk measure, $P(E_j|\boldsymbol{f})$, can be expressed simply by using a standard conditional expectation relationship as

$$\begin{aligned} P(E_j|\boldsymbol{f}) &= E[P(E_j|\boldsymbol{F})|\boldsymbol{f}] \\ &= E\left[1/F_j|\boldsymbol{f}\right] \end{aligned} \tag{4}$$

where the expectation is with respect to the posterior distribution $P(\boldsymbol{F}|\boldsymbol{f})$, evaluated as in (2). Hence, the predictive probability of disclosure event E_j, given knowledge only of sample data $\boldsymbol{f}$ is equal to the posterior mean of $1/F_j$, the reciprocal of the relevant population cell count.

In the model of Benedetti and Franconi (1998), subsequently extended by Rinott (2003) and Polettini and Stander (2004), the posterior distribution (2) simplifies to

$$P(\boldsymbol{F}|\boldsymbol{f}) = \prod_j P(F_j|f_j), \tag{5}$$

where each $P(F_j|f_j)$ is a negative binomial probability function in Benedetti and Franconi (1998) and Rinott (2003), and a more complex expression in Polettini and

Stander (2004). In (5), not only are the population cell frequencies conditionally independent given the sample cell frequencies, as pointed out by Rinott (2003), but perhaps more notably the posterior distribution for F_j given $\boldsymbol{f}$ can be written as $P(F_j|f_j)$, and so F_j is also independent of the *sample* cell frequencies in all other cells. In other words, for estimating the disclosure risk in cell j, the only pertinent information is the sample frequency in that cell.

Where *empirical* Bayes estimation is used, as suggested by Rinott (2003) following Bethlehem et al (1990), the observation above is no longer strictly true, as $P(F_j|f_j)$ is replaced by $\hat{P}(F_j|f_j)$, where the maximisation is performed over the parameters of the prior distribution $P(\boldsymbol{F})$ in (2). This quantity *does* now typically depend on cell frequencies other than f_j. However, for the models which have been typically proposed, it does so in a way which is completely invariant to any permutation of the cell frequencies in other cells. In other words, all that is relevant are the sizes of the cell frequencies in the other cells, and not their positions in the table. The tabular structure of the data is completely ignorable.

Skinner and Holmes (1998) and Elamir and Skinner (2004) adapt the original model of Bethlehem et al (1990) in a way which respects the table structure and hence allows more of the information in the data to be incorporated into disclosure risk estimation. Their approach is equivalent to proposing a prior distribution $P(\boldsymbol{F})$ in (2) which is based on a *log-linear model* for the underlying contingency table. The parameters of the log-linear model are then estimated in empirical Bayes fashion. The sensitivity surrounding which log-linear model to use is somewhat averted by choosing a relatively simple model, but allowing some divergence from the model.

In this paper, we follow, and develop, the approach of Forster and Webb (2005). This approach is also based on log-linear models for the contingency table of population cell frequencies, but the requirement to choose a model *a priori* is avoided, and any model uncertainty is coherently incorporated into the resulting inferences. The approach is described in detail in the next section.

2 A Bayesian model

Following Omori (1999), we assume that $\boldsymbol{F}$ has a multinomial$(n, \boldsymbol{\pi})$ prior distribution. Then, we assume a log-linear model for $\boldsymbol{\pi}$. In the current paper, we shall restrict consideration to those log-linear models which are *decomposable graphical models*. For a broader class of log-linear models, see Forster and Webb (2005). The advantage of considering only decomposable graphical models is that computation is made significantly more tractable and efficient. There is some loss of model flexibility, but the decomposable graphical models still constitute a highly flexible model class. For further details of decomposable graphical models, see Lauritzen (1996). For a decomposable graphical model, we shall write

$$\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta}_m) \tag{6}$$

where m indexes the particular model under consideration, and $\boldsymbol{\beta}_m$ is the corresponding vector of model parameters, which is of lower dimension than $\boldsymbol{\pi}$. For a decomposable model, $\boldsymbol{\beta}_m$ may be considered to be a collection of marginal probabilities corresponding to those subtables of the full contingency table which are unconstrained by the model. As a prior distribution for $\boldsymbol{\beta}_m$, we use the hyper-Dirichlet family, a class of prior distributions based on the Dirichlet distribution for the saturated model (no log-linear constraints) and developed by Dawid and Lauritzen (1994). A hyper-Drichlet distribution consists of a Dirichlet distribution on each set of marginal cell probabilities (sub-vector of $\boldsymbol{\beta}_m$) which are unconstrained by the model. These marginal Dirichlet distributions are dependent, where they have common margins. For example, consider a three-way table with cross-classifiying variables A, B and C. The model AB+BC does not constrain the AB or BC marginal distributions, so the corresponding hyper-Dirichlet distribution is composed of Dirichlet distributions for these two margins, but constrained to give a common set of probabilities for the marginal distribution of B.

Having specified a prior distribution, and observed sample cell frequencies $\boldsymbol{f}$, inference concerning disclosure risk is obtained from the posterior distribution $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f})$. Assuming a sampling scheme under which records are exchangeable, for example Bernoulli sampling or simple random sampling without replacement (the model can be adapted when this is not appropriate), then (3) can be written as

$$P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f}) = \int P(\boldsymbol{F}-\boldsymbol{f}|N-n,\boldsymbol{\beta}_m)P(\boldsymbol{\beta}_m|\boldsymbol{f})\mathrm{d}\boldsymbol{\beta}_m. \tag{7}$$

The first term in the integrand, $P(\boldsymbol{F}-\boldsymbol{f}|N-n,\boldsymbol{\beta}_m)$, is simply a multinomial probability function, with multinomial probabilities $\boldsymbol{\pi}$ determined from $\boldsymbol{\beta}_m$ using (6). Using Bayes theorem, the second term of the integrand is

$$P(\boldsymbol{\beta}_m|\boldsymbol{f}) \propto P(\boldsymbol{f}|n,\boldsymbol{\beta}_m)P(\boldsymbol{\beta}_m),$$

the product of a multinomial probability function for $\boldsymbol{f}$ and the prior density for $\boldsymbol{\beta}_m$.

To this point, we have only described inference under a single log-linear model. In practice, it is unlikely that we will be certain about which model is the most appropriate for building the prior distribution for $\boldsymbol{F}$. A Bayesian approach allows this uncertainty to be coherently incorporated into the prior distribution. Let M denote the set of possible models, and suppose that prior uncertainty about m is encapsulated by a prior distribution over M, involving a set of prior model probabilities $P(m)$. In practice, a discrete uniform distribution over M is commonly used, to represent prior ignorance. The prior distribution over $\boldsymbol{F}, m$ and $\{\boldsymbol{\beta}_m, m \in M\}$ now consists of three components, the multinomial $P(\boldsymbol{F}|\boldsymbol{\beta}_m, m)$, the prior for the parameters of each possible decomposable graphical model $P(\boldsymbol{\beta}_m|m)$ and the prior model probabilities $P(m)$. Note that the first two distributions are now explicitly conditional on m, as both the form of the log-linear model in (6), and the prior distribution for its parameters, will depend on which model is under consideration.

Under model uncertainty, the posterior distribution for the unobserved cell counts $\boldsymbol{F}-\boldsymbol{f}$ in (7) becomes

$$P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f})=\sum_{m\in M}P(m|\boldsymbol{f})\int P(\boldsymbol{F}-\boldsymbol{f}|N-n,\boldsymbol{\beta}_m,m)P(\boldsymbol{\beta}_m|\boldsymbol{f},m)\mathrm{d}\boldsymbol{\beta}_m. \tag{8}$$

The posterior model probabilities, which appear in (8) but not (7) are obtained, using Bayes theorem as

$$P(m|\boldsymbol{f})=\frac{P(m)P(\boldsymbol{f}|m)}{\sum_{m\in M}P(m)P(\boldsymbol{f}|m)} \tag{9}$$

where $P(\boldsymbol{f}|m)$ is the *marginal likelihood* for the sampled cell counts, obtained as

$$P(\boldsymbol{f}|m)=\int P(\boldsymbol{f}|m,\boldsymbol{\beta}_m)P(\boldsymbol{\beta}_m|m)\mathrm{d}\boldsymbol{\beta}_m. \tag{10}$$

The posterior distribution (8) under model uncertainty is obtained as a weighted average of the posterior distributions (7) under the various models. This is sometimes referred to as model-averaging. Care is required when performing model-averaging, that the quantity which is being averaged is one which shares a common interpretation across the component models. That is clearly the case here, where we are averaging probabilities for cell frequencies. The posterior model probabilities are not of interest in themselves, as we do not actually believe that the population was generated by a particular multinomial log-linear model. Their function is to indicate the appropriate weight, based on the sample data, to be applied to the various models in any inference required. Consequently, they determine the differential impact of other cells, when making inference about a particular population cell frequency.

Having obtained the posterior distribution of the unobserved cell frequencies $\boldsymbol{F}-\boldsymbol{f}$ as in (8), it simply remains to evaluate the risk measure (4), using

$$E[1/F_i|\boldsymbol{f}]=\sum_{i=0}^{N-n}\frac{1}{f_j+i}P(F_j-f_j=i|\boldsymbol{f}) \tag{11}$$

where $P(F_j-f_j=i|\boldsymbol{f})$ is the marginal posterior probability obtained from (8) by

$$P(F_j-f_j=i|\boldsymbol{f})=\sum_{\mathbf{F}-\mathbf{f}:F_j-f_j=i}P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f}). \tag{12}$$

3 Computation

There are three computational difficulties associated with calculating the predictive probabilities which are proposed as disclosure risk measures. The first is the evaluation of the integrals in (8) and (10). These integrals are analytically intractable for general log-linear models, but can be straightforwardly evaluated when a hyper-Dirichlet prior distribution is used for a decomposable graphical model. The second problem is evaluation of the sum in (8), in cases where the number of models is so large that evaluation of the summand for every model is infeasible. For example, for a six-way contingency table, as in our second example below, there are many thousands of possible decomposable graphical models. Finally, evaluation of the sum in (12) can also be impracticable, as it involves summing over the sample space for all cells except the one currently of interest.

For decomposable models and hyper-Dirichlet prior distributions, some of the calculations can be performed exactly. For example, provided that the number of models under consideration is not too great, marginal likelihoods (10) are available as ratios of products of gamma functions, and hence (9) may be evaluated directly. For more than a few (3 or 4) cross-classifying variables, it is unlikely to be feasible to calculate posterior probabilities for all models. In such examples, Forster and Webb (2005) use a Markov chain Monte Carlo (MCMC) approach to sampling, suggested by Madigan and York (1996). Here, we rather use an efficient search stategy for identifying a subset of posterior models with high probability, and estimate posterior model probabilities for this set using (9), assuming $P(m|\boldsymbol{f}) = 0$ for all m not in our candidate set. Our search strategy is based on the 'Occam's window' approach of Madigan and Raftery (1994).

Having obtained model probabilities, we are required to evaluate (8) and hence (12). The integral in (8) is tractable for decomposable graphical models and hyper-Dirichlet prior distributions. However, the sheer size of the sample space for $\boldsymbol{F}-\boldsymbol{f}$ in practical disclosure risk assessment problems makes complete enumeration infeasible. An alternative is to replace $\boldsymbol{F} - \boldsymbol{f}$ in (8) with $F_j - f_j$, and hence obtain $P(F_j - f_j|\boldsymbol{f})$ directly. However, when $P(\boldsymbol{F} - \boldsymbol{f}|N - n, \boldsymbol{\beta}_m, m)$ is replaced by the binomial probability $P(F_j - f_j|N - n, \boldsymbol{\beta}_m, m)$ in the integrand of (8), the integral is no longer tractable. These calculations are, however, easily approximated by Monte Carlo sampling from the predctive distribution $P(F_j - f_j|\boldsymbol{f})$. This is easily achieved, and just requires sampling from various Dirichlet and binomial distributions. Then, the probabilities $P(F_j - f_j|\boldsymbol{f})$ in (12) are simply estimated by sample proportions, which can then be plugged into (11), avoiding the requirement to evaluate (12) by summation. An alternative 'Rao-Blackwellized' calculation described by Forster and Webb (2005) avoids any binomial sampling and reduces Monte Carlo error.

4 Examples

We present two examples. The first is a small example, to illustrate the methodology. The second example is more realistic in terms of size and complexity and is presented to illustrate that the methodology is practicable in disclosure risk assessment applications.

4.1 Example 1: A three-way table

To illustrate the model-averaging approach, we consider the data used by Fienberg and Makov (1998) to illustrate their approach. It is a three-way table representing cross-classification by gender, race and income for a selected US census tract.

		Gender					
		Male			Female		
		Income			Income		
		1	2	3	1	2	3
	White	96	72	161	186	127	51
Race	Black	10	7	6	11	7	3
	Chinese	1	1	2	0	1	0

Table 1: Three-way table from Fienberg and Makov (1998). Income categories are 1. ≤\$10000, 2. >\$10000 and ≤\$25000, 3. >\$25000

Interest focusses on the three uniques in this sample, which we label (C,M,1), (C,M,2) and (C,F,3). Following Fienberg and Makov (1998) we shall investigate the potential disclosure risk of this release, in the cases that the sample represents 10%, 20% and 50% of the population.

We consider two possible sets of hyper-Dirichlet prior parameters, both of which have been suggested as implying weak information concerning the model parameters. In both cases all marginal priors are derived from a symmetric Dirichlet distribution (all parameters taking common value α) for the saturated model. For the first prior, $\alpha = 1/2$, and for the second prior $\alpha = 1/K$, the reciprocal of the number of cells in the table (here, $K = 18$). Where the number of cells is large, then the second prior is likely to be preferred, as $K\alpha$ is a measure of the information in the prior, which increases with K in the first case, but is fixed at 1 in the second case.

Table 3 presents the measure of disclosure risk $E[1/F_j|\boldsymbol{f}]$ for the three sample unique cells, for both priors and all three sampling fractions. The first thing to notice is that the inferences for the three cells are different, although not dramatically so. Hence the approach is having the desired effect of incorporating information about the structure in the table. For both priors, the posterior distribution is concentrated on a small selection of models. For prior 1, models R+IG and RG+IG dominate, while for prior 2, model R+IG dominates.

		Prior 1 ($\alpha = 1/2$)			Prior 2 ($\alpha = 1/18$)		
		Sampling fraction			Sampling fraction		
		50%	20%	10%	50%	20%	10%
	(C,M,1)	0.595	0.235	0.109	0.707	0.344	0.173
Cell	(C,M,2)	0.665	0.295	0.142	0.766	0.422	0.226
	(C,F,3)	0.570	0.221	0.105	0.651	0.293	0.143

Table 2: Monte Carlo Estimates of $E[1/F_j|\boldsymbol{f}]$

4.2 Example 2: A six-way table

To test the methodology on a more realistic example, we extracted a six-way table of potential key variables from the 3% Individual Sample of Anonymized Records (SAR) for the 2001 UK Census (see `http://www.ccsr.ac.uk/sars/2001` for full details). The table extracted consisted of 154295 individuals living in South West England, cross-classified by sex (2 categories) age (coded into 11 categories), accomodation type (8 categories), number of cars owned or available for use (5 categories), occupation type (11 categories) and family type (10 categories). The full table has 96800 cells of which 3796 are uniques. For the purposes of this exercise, this is considered to be the population. To mimic the selection into the SAR, we took a 3% subsample, containing 4761 individuals of whom 1543 were uniques. In the sample data, only 2330 of the 96800 cells were non-empty. Hence, 32% of records, and 66% of cells correspond to sample uniques. Of these cells, only 114 (7%) are population uniques, and the average population total in a sample unique cell is 17, so not all such cells represent disclosure risk.

For each of the 2330 non-empty cells j, we calculated the predictive disclosure probability $P(E_j|\boldsymbol{f}) \equiv E[1/F_j|\boldsymbol{f}]$. For this exercise, we are also able to calculate the probability of a disclosure event $P(E_j|\boldsymbol{F}) \equiv 1/F_j$ in the case that full population knowledge was available. We compare these quantities, and hence assess the performance of our disclosure risk assessment procedure by plotting $\log_{10}(1/F_j)$ against the estimated $\log_{10}(E[1/F_j|\boldsymbol{f}])$ for the 2330 non-empty sample cells, in Figure 1.

Given that these are the lowest frequency cells in the table, accurate estimation is a difficult task, and the model-averaging approach seems to be performing quite well, with perhaps a slight tendency to overestimate risk in this example. For comparison, we note that without any log-linear modelling, the estimated risk $E[1/F_j|\boldsymbol{f}]$ for any sample unique is evaluated as 0.11, so it is immediately clear that our approach is providing a more accurate measure of risk for the cells with low population counts (genuinely risky records). Indeed, for the 114 genuine population uniques, we computed an average risk of 0.65, while for the 111 sample unique cells with population totals greater than 50, the average risk was estimated as only 0.04. So the method is demonstrating an ability to distinguish risky and non-risky cells with the same cell counts. One way of assessing the performance of the method is by considering it as a classifier. Suppose that we define a cell as 'risky' if the probability of a disclosure event is greater than 5%. Then, our method classifies cells as risky if $(E[1/F_j|\boldsymbol{f}]) > 0.05$. The 'true' classification is determined using the corresponding (unobserved) value of $1/F_j$. Table 3, shows how our classifier performs. In these terms, the performance seems quite satisfactory, given the small sampling fraction. The sensitivity of the classifier is 88% and its specificity is 76%.

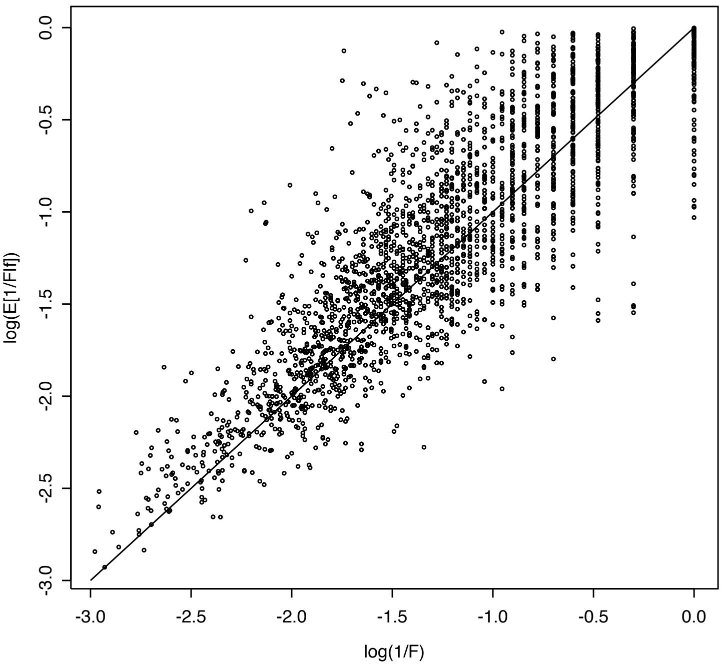

Figure 1: The estimated $\log_{10}(E[1/F_j|\boldsymbol{f}])$ against $\log_{10}(1/F_j)$ for the 2330 non-empty sample cells. The plotted line represents equality (no error).

		True classification	
		Not risky	Risky
Estimated classification	Not risky	864	140
	Risky	267	1059

Table 3: Performance of model averaging as a risk classifier, assessed using the 2330 non-empty cells of the sample data table.

5 Discussion

The examples presented in Section 4 illustrate that this approach has potential for identifying cells which may pose a disclosure risk. With the second example, we have started to investigate the performance of the methodology on more realistic examples. In fact, the computational time for this example was not large. It took a few seconds to compute using functions written in R. There therefore remains scope to extend to much more demanding examples.

References

Benedetti, R and Franconi, L. (1998). Statistical and technical solutions for controlled data dissemination. In *Pre-Proceedings of New Techniques and Technologies for Statistics, Volume 1* 225–232. Sorrento, Italy.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21**, 1272–1317.

Elamir, E. A. H. and Skinner, C. J. (2004). Record-level measures of disclosure risk for survey microdata. *S^3RI Methodology Working Paper*, **M04/02**. Southampton Statistical Sciences Research Institute.

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`.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

Madigan, D. and Raftery, A. E. (1995). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

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. and Stander, J. (2004). A hierarchical Bayesian model approach to risk estimation in statistical disclosure limitation. In *Privacy in Statistical Databases*, J Domingo-Ferrer and V Torra (Eds), 247–261. Springer Lecture Notes in Computer Science, 3050, 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.

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