WP. 10
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

A NEIGHBORHOOD REGRESSION MODEL FOR SAMPLE DISCLOSURE RISK ESTIMATION

Invited Paper

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

[1] Prepared by Yosef Rinott, Hebrew University, Israel and Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton.

A neighborhood regression model for sample disclosure risk estimation.

Yosef Rinott*, Natalie Shlomo**

* Department of Statistics, Hebrew University, Jerusalem, Israel. (rinott@mscc.huji.ac.il)

** Department of Statistics, Hebrew University of Jerusalem, Southampton Statistical Sciences Research Institute, University of Southampton, UK. (N.Shlomo@soton.ac.uk)

Abstract. The disclosure risk involved in releasing data which consist of a sample from some population depends on both the sample and the population. When the sample is fully known, with only partial or no information on the population, a major problem in *Statistical Disclosure Control* (SDC) is the estimation of *disclosure risk* on the basis of the sample. Considering data in the form of a frequency table, risk arises from non-empty sample cells which represent small population cells (and population uniques in particular). Therefore risk estimation requires assessing which of the relevant population cells are indeed small.

Various methods have been proposed for this task, and we present a new one, in which estimation of population cell frequencies is based on a model connecting the table parameters in neighborhoods defined in natural ways using the table structure and the nature of the variables. At this point this method is under experimentation, and we provide some preliminary comparisons with the *Argus method* in which inference is based on sampling weights, and with a *log-linear models* approach.

1 Introduction

Let $\mathbf{f} = \{f_k\}$ denote an m-way sample frequency table, where $k = (k_1, ..., k_m)$ indicates a cell and f_k is the frequency in the cell, and let $\mathbf{F} = \{F_k\}$ denote the population from which the sample is drawn. We denote the sample and population sizes by n and N respectively, and the number of cells by K. Disclosure risk arises from cells in which both f_k and F_k are positive and small, and in particular when $f_k = F_k = 1$ (a sample and population unique).

Various individual and global risk measures have been proposed in the literature, see e.g., Benedetti, Capobianchi and Franconi (1998), Skinner and Holmes (1998), Elamir and Skinner (2006), Rinott (2003). In this paper we chose to focus only on two global risk measures,

$$\tau_1 = \sum_k \mathbb{I}(f_k = 1, F_k = 1)\,, \qquad \tau_2 = \sum_k \mathbb{I}(f_k = 1)\frac{1}{F_k}$$

where $\mathbb{I}$ denotes the indicator function. Note that τ_1 counts the number of *sample uniques* which are also *population uniques*, and τ_2 is the expected number of correct

guesses if each sample unique is matched to a randomly chosen individual from the same population cell. These measures are somewhat arbitrary, and one could consider measures which reflect matching of individuals that are not sample uniques, possibly with some restrictions on cell sizes. Also, it may make sense to normalize these measures by some measure of the total size of the table, by the number of sample uniques, or by some measure of the information value of the data. Such and other measures should also be considered.

When only $\mathbf{f}$ is known, and $\mathbf{F}$ is considered an unknown parameter (on which there is often some partial information) the quantities τ_1 and τ_2 should be estimated. Note that they are not proper parameters, since they involve both the sample $\mathbf{f}$ and the parameter $\mathbf{F}$. Therefore a discussion of the variances of estimates of τ_1 and τ_2 requires special care, see Rinott (2003) for some details, and Zhang (2005) for general theory. We shall discuss this issue in a subsequent paper.

In this paper we describe two known methods of estimation of quantities like τ_1 and τ_2, propose a new one, and compare them by some experiments. The first by Benedetti, Capobianchi and Franconi (1998) which uses the *Negative Binomial* model, provides the basis to the *μ-Argus* program, and the second, proposed by Skinner and Holmes (1998) and Elamir and Skinner (2006), uses a *Poisson* model and bases estimation on hierarchical *log-linear models*. The new method we propose is based on a different model which we shall explain. We shall present here the main ideas of this method, which is under development, and preliminary experiments .

All the above methods consist of modeling the conditional distribution of $\mathbf{F}|\mathbf{f}$, estimating parameters in this distribution and then using estimates of the form

$$\hat{\tau}_1 = \sum_k \mathbb{I}(f_k = 1)\hat{P}(F_k = 1|f_k = 1), \qquad \hat{\tau}_2 = \sum_k \mathbb{I}(f_k = 1)\hat{E}[\frac{1}{F_k}|f_k = 1] \quad (1)$$

where $\hat{P}$ and $\hat{E}$ denote estimates of the relevant conditional probability and expectation. For a general theory of estimates of this type see Zhang (2005) and reference therein.

2 Models

For completeness we briefly introduce the Poisson and Negative Binomial models. More details can be found, for example, in Bethlehem et al (1990), Cameron and Trivedi (1998), Rinott (2003).

A common assumption in the frequency table literature is $F_k \sim$ Poisson$(N\gamma_k)$, independently, with $\sum \gamma_k = 1$. Binomial (or Poisson) sampling from F_k means that $f_k|F_k \sim Bin(F_k, \pi_k)$, π_k being the sampling fraction in cell k. By standard calculations we then have

$$f_k \sim \text{ Poisson}(N\gamma_k\pi_k) \text{ and } F_k \,|\, f_k \;\sim\; f_k + \text{ Poisson}(N\gamma_k(1-\pi_k))\,, \qquad (2)$$

leading to the Poisson model of subsection 2.1 below.

If one adds the Bayesian assumption $\gamma_k \sim$ Gamma(α, β) independently, with $\alpha\beta = 1/K$ to ensure that $E\sum\gamma_k = 1$, then $f_k \sim NB(\alpha, p_k = \frac{1}{1+N\pi_k\beta})$, the Negative

Binomial distribution defined for any $\alpha > 0$ by $P(f_k = x) = \frac{\Gamma(x+\alpha)}{\Gamma(x)\Gamma(\alpha)}(1-p_k)^x p_k^\alpha$, $x = 0, 1, 2, \ldots$, which for a natural α counts the number of *failures* until α successes occur in independent Bernoulli trials with probability of success p_k. Further calculations yield $F_k \,|\, f_k \sim f_k + NB(\alpha + f_k, \frac{N\pi_k + 1/\beta}{N + 1/\beta})$, $(F_k \geq f_k)$.

As $\alpha \to 0$ (and hence $\beta \to \infty$) we obtain $F_k \,|\, f_k \sim f_k + NB(f_k, \pi_k)$, which is exactly the Negative Binomial assumption in Section 2.2 below. As $\alpha \to \infty$ the Poisson model of Section 2.1 is obtained, and in this sense the Negative Binomial with $\alpha \neq 0$ subsumes both models. Applications of this generalization will be given in a subsequent paper.

2.1 The Poisson log-linear method

Skinner and Holmes (1998) and Elamir and Skinner (2006) proposed and studied the following approach. Assuming a fixed sampling fraction, that is, $\pi_k = \pi$, the first part of (2) implies $f_k \sim$ Poisson$(n\gamma_k)$, where $n = N\pi$. Using the sample $\{f_k\}$ one can fit a log-linear model using standard programs, and obtain estimates $\{\hat{\gamma}_k\}$ of the parameters. Using the second part of (2) it is easy to compute

$$P(F_k = 1 | f_k = 1) = e^{-N\gamma_k(1-\pi_k)}, \; E[\frac{1}{F_k} | f_k = 1] = \frac{1}{N\gamma_k(1-\pi_k)}[1 - e^{-N\gamma_k(1-\pi_k)}]. \quad (3)$$

Plugging $\hat{\gamma}_k$ for γ_k in (3) leads to the desired estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1). The quantity $E[\frac{1}{F_k} | f_k = 1]$ is sometimes referred to as the *individual risk measure* at cell k.

2.2 The Negative Binomial Argus method

In this method, proposed by Benedetti, Capobianchi and Franconi (1998), see also Polettini and Seri (2003), it is assumed that $F_k | f_k \sim f_k + NB(f_k, \pi_k)$. There is an implicit assumption of independence between cells.

Using the relation $E_{\pi_k}[F_k | f_k] = f_k/\pi_k$, the parameters π_k are estimated using sampling weights: if w_i denotes the sampling weight of individual i, then an *initial estimate* of F_k is $\hat{F}_k = \sum_{i \in \text{ cell } k} w_i$, and we obtain the moment-type estimate $\hat{\pi}_k = f_k/\hat{F}_k$. Straightforward calculations with the Negative Binomial distributions show

$$P_{\hat{\pi}_k}(F_k = 1 | f_k = 1) = \hat{\pi}_k, \quad \text{and} \quad E_{\hat{\pi}_k}[\frac{1}{F_k} | f_k = 1] = -\frac{\hat{\pi}_k}{1 - \hat{\pi}_k} \log(\hat{\pi}_k) \, .$$

Plugging these estimates for $\hat{P}$ and $\hat{E}$ in (1) we obtain the estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ of the global risk measures. Note that in this method the cells are treated completely independently, each cell at a time, and the structure of the table plays no role.

2.3 A brief discussion

Estimation of risk measures without a model which restricts the number of parameters, such as a log-linear model, is inherently difficult. To see this just note that if one estimates γ_k in each cell separately without a model by $\hat{\gamma}_k = f_k/n$ then the estimated population cell frequency $N\hat{\gamma}_k$ satisfies $\text{Var} N\hat{\gamma}_k \approx N^2\gamma_k/n$. Typically, risk

arises from cells where $\gamma_k = O(1/N)$ since such cells are likely to contain population uniques, and for such k we obtain $\mathrm{SD}(N\hat{\gamma}_k) = O((N/n)^{1/2})$ which is usually large.

The situation improves in the presence of a model that reduces the number of parameters, provided of course that the model is valid. In order to see this in a specific example, consider a two-way table and the (log-linear) model of independence. For the Maximum Likelihood estimate $\hat{\gamma}_k$, it can then be shown directly that the variances of the cell frequency estimate $N\hat{\gamma}_k$, or that of $\hat{P}(F_k = 1|f_k = 1) = e^{-N\hat{\gamma}_k(1-\pi_k)}$ which appears in the estimate $\hat{\tau}_1$ is $O(\frac{N^2}{n^2}\gamma_k + \frac{N^2}{n}\gamma_k^{1+\nu})$ for some $\nu \leq 1/2$ which depends on the parameters. Again looking at cells where $\gamma_k = O(1/N)$, and $\nu = 1/2$, the standard deviation $\mathrm{SD}(N\hat{\gamma}_k)$ of the cell frequency estimate is like $O(N^{1/2}/n)$, a great improvement, and often small enough. The situation improves further with large higher order tables if simple models, like independence, are valid (see Zhang 2005). The above also shows that dividing the population into smaller parts will increase the variance, and should be done only if it leads to better models.

The estimation question here is essentially the following: given, say, a sample unique, how likely is it to be also a population unique, or arise from a small population cell. The *Argus* method bases its estimation on sampling weights (and the NB model). There is no learning from other cells. However, such learning appears natural. If a sample unique is found in a part of the table where neighboring cells (by some reasonable metric, to be discussed later) are small or empty, then it seems reasonable to believe that it is more likely to have arisen from a small population cell.

As we saw above, when a *log-linear model* is indeed valid, it will reduce the standard deviation of population frequency estimates and hence of risk measures. The *log-linear model* approach indeed uses cells from neighborhoods which depend on the model to determine the risk in a given cell. For example, if the attributes forming the table are assumed independent, then the estimate $\hat{\gamma}_k$ is the product of marginals obtained by fixing one attribute at a time, so that every cell which has a common value with one of the attributes of cell k will contribute to the risk estimate at this cell; thus if one of the attributes is economic status, then inference on the very rich involves also information from the very poor, provided they have some other attribute in common, such as marital status.

This observation led us to trying another type of *neighborhoods*, thinking that log-linear models, which provide explanations to the data when they fit, may not lead to the most natural neighborhoods for the question at hand. Our initial attempts will be described in Section 2.4 and some experiments are described in Section 3.

Another inherent problem that arises in the *Argus* method is related to the fact that for empty sample cells the *initial estimates* $\hat{F}_k$ vanish. Since the total population estimate should be N, it follows that other population cells tend to be overestimated, and as a result, risk measures are underestimated. A systematic treatment of this hard problem would require identifying *structural zeros* and perhaps replacing other sample zeros by some ε, as sometimes proposed in the literature in the context of model building. It is easy to see that if this is introduced into the *Argus* method, risk measures will increase to the correct values as ε increases, and then will exceed them. However, estimation of the right ε, or similar parameters, appears difficult.

A version of the latter issue appears also in the *log-linear model* approach. If a saturated model is used then $\hat{\gamma}_k = f_k/n$, and for empty sample cells $\hat{\gamma}_k = 0$, leading to

the same underestimation problem as above. In fact, for $f_k = 1$ we have $N\hat{\gamma}_k = N/n$ and from (3) we obtain $\hat{P}(F_k = 1|f_k = 1) \approx e^{-N/n}$ and $\hat{E}[\frac{1}{F_k}|f_k = 1] \approx n/N$, so that all sample uniques are estimated to have the same very low risk. At the other extreme, if we take a model of independence then $\hat{\gamma}_k$ is obtained as a product of terms where each term is a large sum of frequencies over all attributes except for one, that is, for $k = (k_1, ..., k_m)$, $\hat{\gamma}_k = \prod_i (\sum_{k_j, j \neq i} f_k/n)$. The large sum for a given i in the latter product vanishes only if the level k_i of the attribute i never appears in the sample, and in that case it would probably be omitted from the file. Thus, the model usually has no zero population cell predictions, and in view of the above one should expect higher risk estimates. In fact the independence model often leads to overestimation of risk as expected by this explanation. Intermediate models, such as those of conditional independence involve products of smaller sums, and in general one may expect monotonicity of the risk estimates in the size of the model (number of parameters). So again, as in the choice of ε above, there is usually a model which would give a good risk estimate for a given risk measure. The question of finding goodness of fit measures so that the model chosen provides good risk estimates is studied in Skinner and Shlomo (2005).

2.4 Neighborhoods

We consider frequency tables in which some of the attributes are ordinal. For such an attribute i we can consider a set of levels S_{k_i} which are close to a given level k_i in the attribute's ordering. Given cell $k = (k_1, ..., k_m)$ we can construct a neighborhood of cells N_k of k, by varying the coordinates k_i of the ordinal attributes in some way in the sets S_{k_i}, and fixing the other, non-ordinal attributes.

More specifically, let O denote the set of ordinal attributes and suppose the attribute $i \in O$ has levels $1, 2, \ldots, r_i$. Here we consider neighborhoods of the type $N_c^k = \{h = (h_1, ..., h_m) : \sum_{j \in O} |h_j - k_j| = c,\ h_j = k_j \text{ for } j \notin O\}$ or the type $M_a^k = \{h = (h_1, ..., h_m) : |h_j - k_j| = a_j \text{ for } j \in O,\ h_j = k_j \text{ for } j \notin O\}$ for some $a = (a_1, ..., a_m)$.

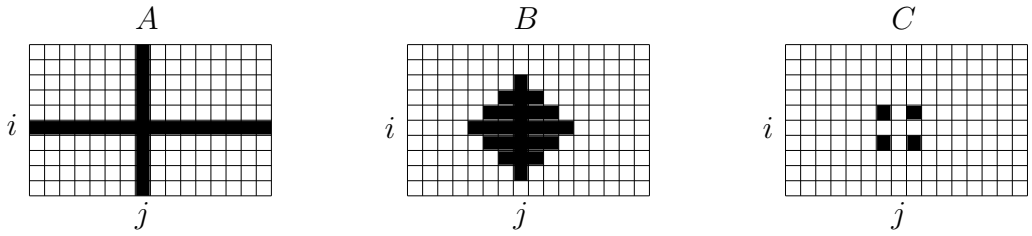

Figure 1. Neighborhood of cell $k = (i, j)$. A: under independence model. B: the union of neighborhoods $\bigcup_{c \leq 3} N_c^k = \bigcup_{|a| \leq 3} M_a^k$. C: the neighborhood $M_{(1,1)}^k$

This approach can perhaps be extended to non-ordinal attributes having some metric or a measure of proximity between their levels, such as geographic location.

The neighborhoods are used as follows: we assume as in (2), with $\pi_k = \pi$ for simplicity, $f_k \sim$ Poisson$(N\gamma_k\pi)$ and $F_k \,|\, f_k \sim f_k +$ Poisson$(N\gamma_k(1-\pi))$, but now we propose to consider log-linear models of the form $\gamma_k = exp\{\beta_0 + \sum_{c \leq C} \beta_c x_c^k\}$ for some C to be determined, where $x_c^k = \sum_{\ell \in N_c^k} f_\ell$. We can estimate the parameter vector β, obtain estimates $\hat{\gamma}_k$ and proceed to estimate risk as before using these

estimates in the above Poisson conditional distribution of $F_k \mid f_k$. In a similar way, setting $|a| = \sum_{j \in O} |a_j|$, we tried this model with $\gamma_k = exp\{\beta_0 + \sum_{a:\,|a| \le C} \beta_a z_a^k\}$ for some C to be determined, where $z_a^k = \sum_{\ell \in M_a^k} f_\ell$.

Other regression models (e.g., Negative Binomial, see Cameron and Trivedi (1998) for Poisson and Negative Binomial regression) and types of neighborhoods, and combinations of the neighborhood approach with weights and other information on the population will be discussed in a subsequent paper.

Regarding the issue of *structural zeros*, we tried declaring a cell to be a structural zero if all its neighborhoods which are used in the regression contain only empty cells.

Some technical issues: The cardinality of N_c^k satisfies $|N_c^k| = 2^m \sum_{t=\min(m-c,\,0)}^{m-1} 2^{-t} \binom{m}{t} \binom{c-1}{m-t-1}$ which increases rapidly with m and c (it is smaller for k's near the boundary of the table, but still many of these neighborhoods are rather large). For $m = 4$ (a four-way table) we have for k's not near the boundary $|N_5^k| = 360$ and $|N_7^k| = 856$. On the other hand, the neighborhoods M_a^k are not as large, however, the number of neighborhoods of the type M_a^k with $|a| = c$ is $\binom{c+m-1}{c}$, so that for $m = 4$ and $c = 7$, for example, we would have to deal with 120 such neighborhoods and β coefficients in the regression. Therefore our preliminary experiments presented in the Section 3 are quite restricted in size and perhaps not very impressive at this point. There is much room for improving and fine-tuning the method and the programs, and for testing different types of data before conclusions can be drawn.

3 Experiments with neighborhoods

We present a few experiments. They are preliminary as already mentioned and more work is needed on the approach itself and on classifying types of data for which it might work.

In the experiments we used our versions of the Argus and log-linear models approaches, programmed on the SAS system. In all experiments we took a real population data file of size N given in the form of a contingency table with K cells, and from it we took a random sample of size n. Since the population and the sample are known to us, we can compute the *true values* of τ_1 and τ_2 and their estimates by the different methods, and compare.

Example 1 In this example the population consists of an extract of the 1995 Israeli Census Sample File for Individuals with age 15 and over with $N = 746,949, n = 14,939$, and $K = 337,920$. The attributes (with number of levels in parentheses) were Sex (2), Age Groups (16), Groups of Years of Study (10), Number of Years in Israel (11), Income Groups (12), and Number of Persons in Household (8). Since Sex is not ordinal, neighborhoods were constructed with Sex being fixed and the set of ordinal attributes O contains the other five variables. We used neighborhoods of the type N_c^k for $c \le C = 4$, and M_a^k for $|a| \le C = 4$.

In one version of the experiment we ignored the issue of structural zeros, and in another we define structural zeros as all sample cells that have a zero count and the sum of the sample counts in all of the neighborhoods is zero. Out of $K = 337,920$ cells, we obtained 206,655 non-structural zeros. The Poisson regression model with

the new types of neighborhoods was run on this file with and without the structural zeros to obtain the expected cell means and risk measure estimates as described above. The weights w_i for the *Argus* method in all our examples were computed by post-stratification on Sex by Age by Geographical location (the latter is not one of the attributes in any of the tables, but it was used for post-stratification). These variables are commonly used for post-stratification, other strata may give different results. Two log-linear models are considered, one of independence, the other including all two-way interactions.

Model	τ_1	τ_2
True Values	430	1125.8
Argus	114.5	456
Log Linear Model: Independence	773.8	1774.1
Log Linear Model: 2-Way Interactions	470	1178.1
Neighborhood method M_a^k	786.8	2146.9
Neighborhood method M_a^k excluding structural zeros	385.4	1674.1
Neighborhood method N_c^k	723.3	2099.6
Neighborhood method N_c^k excluding structural zeros	344.8	1624.2

Example 2 This data consist of an extract of the 2001 UK Census file, $N = 944,793$, $n = 18,896$, $K = 152,100$, with the attributes, Sex (2) Age Groups (25) Number of Persons in Household (9) Education Qualifications (13) Occupation (26). Sex was treated as non-ordinal as above.

Model	τ_1	τ_2
True Values	191	568.0
Argus	79.2	315.6
Log Linear Model: Independence	364.8	862.3
Log Linear Model: 2-Way Interactions	182.3	546.2
Neighborhood method M_a^k	42.8	770.0
Neighborhood method M_a^k excluding structural zeros	6.4	540.2
Neighborhood method N_c^k	38.5	755.2
Neighborhood method N_c^k excluding structural zeros	5.6	529.0
Neighborhood method N_c^k with $c \leq 12$	50.6	748.3

Example 3 This example is from the extract of the 1995 Israeli Census Sample File for Individuals aged 15 and over, $N = 248,983$, $n = 2,490$, $K = 8,800$, with attributes Sex (2) Age Groups (16), Years of study (25), and Occupation (11).

Model	τ_1	τ_2
True Values	5	36.9
Argus	7.7	35.5
Log Linear Model: Independence	6.4	44.2
Log Linear Model: 2-Way Interactions	1.1	26.4
Neighborhood method M_a^k	0	30.0
Neighborhood method M_a^k excluding structural zeros	0	25.0
Neighborhood method N_c^k	0	30.1
Neighborhood method N_c^k excluding structural zeros	0	25.5

Discussion of examples In Example 1, the independence log-linear model and the neighborhoods model overestimate the two risk measures. As expected (see Section 2.3), the log-linear model with two-way interactions, which provides the best estimates here, and the exclusion of structural zeros in the neighborhood method yield lower risk estimates. The neighborhood models which take structural zeros into account yield reasonable estimates, while Argus underestimates risk.

In Example 2, again the two-way interaction model wins, while Argus and the neighborhood model with $C = 12$, which requires heavy calculations and therefore was so far done only once, are doing reasonably well.

In Example 3 Argus comes out best, while here the log-linear independence model does well and it is better than the two-way interaction model, which was the winner in the previous two examples, although it is hard to believe that variables like Age, Years of Study, and Occupation can be independent. A similar phenomenon occurred in another experiment from the same file, with the ordinal attributes Age (71, top coded at 85+), Groups of Years of Study (18), and Income Groups (18). The log-linear model of independence gave the best results, although the variables cannot be independent.

This raises the following question: in a multi-way table, how would one choose the right model? Will the best fitting model by standard measures of goodness of fit provide the best risk estimation results? Skinner and Shlomo (2005) deal with this question. In Example 3, the risk estimates from the two-way interaction model are also quite good, but it seems that in higher dimensional tables, with many possible models, the problem of model selection will be crucial.

Our preliminary **conclusions** are that the new neighborhood approach presented here proposes a natural model which like the other methods needs to be refined and fine-tuned. We expect the new model to work well relative to log-linear models in multi-way tables when simple log-linear models are not valid. We intend to incorporate our approach into a more general regression model, the Negative Binomial Regression, which subsumes the Poisson regression model (Cameron and Trivedi 1998), invoke sampling weights and calibration to partial information on the population, and thus combine the new ideas with known aspects of regression models and ideas of Argus. The burden of proof is still on us.

References

Benedetti, R., Capobianchi, A, and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design Information.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure Control of Microdata, *J. Amer. Statist. soc.* , **85**, 38–45.

Cameron, A. C., and Trivedi, P. K. (1998) Regression analysis of count data. *Econometric Society Monographs*, **30**. Cambridge University Press.

Elamir, E. and Skinner, C. (2006) Record-level measures of disclosure risk for survey microdata, *Journal of Official Statistics*, to appear.

Polettini, S. and Seri, G. (2003) Guidelines for the protection of social micro-data using individual risk methodology - Application within mu-argus version 3.2, CASC Project Deliverable No. 1.2-D3, http://neon.vb.cbs.nl/casc/

Polletini, S. and Stander, J. (2004), A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, Springer-Verlag, New York, 247–261

Rinott, Y. (2003) On models for statistical disclosure risk estimation, *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxemburg , 275-285.

Skinner, C. and Shlomo, N. (2005), Assessing disclosure risk in microdata using record-level measures. *In this volume.*

Skinner, C. and Holmes, D. (1998), Estimating the Re-identification Risk Per Record in Microdata, *J. Official Statist.*, **14**, 361-372.

Willenborg, L. and de Waal T. (2001) Elements of Statistical Disclosure Control , *Lecture Notes in Statistics*, **155** , Springer, New York.

Zhang C.-H. (2005) Estimation of sums of random variables: examples and information bounds, to appear in *Ann. Statist.*, **33**.