Journal of Official Statistics, Vol. 22, No. 3, 2006, pp. 525–539

Record Level Measures of Disclosure Risk for Survey Microdata

Elsayed A.H. Elamir[1] *and Chris Skinner*[2]

Measures of disclosure risk at the record level have a variety of potential uses in statistical disclosure control for microdata. We propose a record level measure of disclosure risk, which is the probability that a unique match between a microdata record and a population unit is correct. For discrete key variables subject to no measurement error, we study this measure under the assumption of a Poisson and a Poisson-gamma log-linear model. We apply the approaches to a sample of microdata from the U.K. General Household Survey. The results provide empirical validation of the risk measure.

Key words: Log-linear model; Poisson-gamma model; population uniqueness; statistical disclosure control.

1. Introduction

Researchers require access to survey microdata for analysis, but agencies conducting surveys have obligations to the respondents providing the data and need to protect against statistical disclosure when making microdata available. There is a growing literature on methods for undertaking such protection – see for example, Duncan and Lambert (1989), Bethlehem et al. (1990), Lambert (1993), Fienberg and Makov (1998) and Willenborg and de Waal (2001) and there is an increasing interest in applying these methods in government statistical agencies (Doyle et al. 2001).

In this article we consider the problem of assessing whether a specified form of microdata output could lead to statistical disclosure. Direct identifiers for individuals, such as names and addresses, are assumed to have been removed from the data to form an 'anonymised' file. Disclosure could still arise, however, if the user of the file could identify an individual using the values of the variables recorded in the microdata. We shall use *disclosure risk* as a broad term to refer to the probability of such an event; the precise nature of the event and the probability requires further clarification. The challenge is to construct a measure of disclosure risk, which not only reflects relevant concerns about disclosure, but also can be estimated adequately from the microdata.

Measures of disclosure risk are often based upon the notion of *identifying key variables* (Bethlehem et al. 1990). These are variables with values assumed known both for individuals in the microdata sample and for certain identifiable individuals in the

[1] Management and Marketing Department, College of Business, University of Bahrain, P.O. Box 32038, Bahrain. Email: Sali@buss.uob.bh
[2] Southampton Statistical Sciences Research Institute, University of Southampton, SO17 1BJ, UK. Email: cjs@socsci.soton.ac.uk

population. We shall assume that the relevant units are individuals, but other units, such as households, are possible. An example of a measure of disclosure risk is the proportion of individuals in the microdata sample that have a unique combination of values of the key variables (assumed categorical) in the population (Fienberg and Makov 1998). Such individuals, referred to as *population unique*, may be judged to be particularly 'at risk of disclosure'.

A measure of the form 'the proportion of individuals in the microdata file at risk of disclosure' may be problematic, however, if it is considered unacceptable for disclosure to arise for any individual in the file. In this case, even if one individual out of ten thousand in the microdata sample is seriously 'at risk' then this might be unacceptable, despite the small value (0.0001) of the measure. The basic problem here is that the measure is a "file level" measure, which 'averages the risk' across the whole microdata sample and thus may conceal small parts of the sample where the risk is high.

To address such concerns, it is natural to consider a record level measure, i.e., a measure which may take a different value for each record in the microdata (Elliot 2001). Such a measure may help identify those parts of the sample where disclosure risk is high and more protection is needed and may be aggregated in different ways to a file level measure if desired (Lambert 1993). While record level measures may provide greater flexibility and insight when assessing whether specified forms of microdata output are 'disclosive', they are potentially more difficult to estimate than file level measures.

A number of approaches have been proposed for the estimation of record level measures. For continuous key variables, Fuller (1993) shows how to assess the record level probability of identification in the presence of added noise, under normality assumptions. See also Paass (1987) and Duncan and Lambert (1989). For categorical variables, Skinner and Holmes (1998) define a measure as the probability that a record is population unique, with probability interpreted with respect to a model. They restrict attention to sample unique records, i.e., records with combinations of values of the key variables which are unique in the microdata sample, on the grounds that these are the records most at risk. Like Bethlehem et al. (1990), they assume a compound Poisson model for the generation of the frequencies of the values of the key variables, but with a log-normal distribution for the compound error rather than a gamma distribution. Like Fienberg and Makov (1998), they use a log-linear model to capture the dependence on the key variables. After estimating the model parameters, they use numerical integration to compute the measure. Carlson (2002) develops a related method with the log-normal distribution replaced by an inverse Gaussian distribution.

An alternative approach is currently used in the software μ-ARGUS (Franconi and Polettini 2004), where a measure is defined as the expected value of $1/F_j$, conditional on the observed microdata, where F_j is the population frequency of the record's combination of key values j. Thus, $F_j = 1$ if the record is population unique. Conditional on the data, F_j is assumed to follow a negative binomial distribution.

In this article, we show how the log-linear modelling approach of Skinner and Holmes (1998) may be applied to a risk measure defined as above for μ-ARGUS. We argue that this measure has a useful interpretation as the probability that an observed match (between a microdata record and an identifiable unit in the population) is correct, following an analogous argument for file level measures by Skinner and Elliot (2002). In the μ-ARGUS approach to estimating such a measure, the only variation in the record level measure among records with a common key value sample frequency arises from unequal sample weights. Thus, for example, in a survey with constant weights, all sample uniques will have the same record level measure. In contrast, we show that the log-linear modelling approach can lead to variation in record level measures in this case and we provide empirical validation that this variation measures real differences in disclosure risk. We thus suggest that the proposed approach provides a more realistic measure of record level risk.

A second purpose of this article is to consider the extent to which the log-linear modelling approach of Skinner and Holmes (1998) may be simplified computationally, in particular by replacing the compound Poisson assumption by a simple model without random effects.

The statistical framework for the article is introduced in Section 2. The record level measure is defined in Section 3 and the specification of alternative models and estimation under these models are discussed. The relation of the proposed measure to some file level measures of risk is discussed in Section 4. An empirical evaluation of the approaches outlined in Section 3 is presented in Section 5 based upon data from the UK General Household Survey.

2. Framework and Notation

In this section we introduce the formal framework. We consider a finite population U, consisting of N individuals (or some other form of unit), and suppose that the microdata file consists of records for a sample $s \subset U$ of size $n \leq N$. The sampling fraction is denoted $\pi = n/N$. Following Bethlehem et al. (1990), we assume that the possibility of statistical disclosure arises if an intruder gains access to the microdata and attempts to match a microdata record to external information on a known individual using the values of m discrete key variables $X_1, X_2, \ldots, X_m$.

Let the variable formed by cross-classifying $X_1, X_2, \ldots, X_m$ be denoted X, with values denoted $1, \ldots, J$, where J is the number of categories or key values of X. Each of these key values corresponds to a possible combination of categories of the key variables. As in Section 1, let F_j be the number of units in the population with key value j, i.e., the population frequency or size of cell j for $j = 1, \ldots, J$, and let the population frequencies of frequencies be $N_r = \sum_j I(F_j = r)$, $r = 1, 2, \ldots$. For example, N_1 is the number of population uniques. The sample counterpart of F_j is denoted by f_j and the sample frequencies of frequencies by $n_r = \sum_j I(f_j = r)$, $r = 1, 2, \ldots$. For example, n_1 is the number of sample uniques.

3. Disclosure Risk at the Record Level

3.1. Definition of Risk and Model Specification

We consider a microdata record with key value $X = j$. We restrict attention to records which are sample unique, i.e., $f_j = 1$, since these may be expected to be most risky (Skinner and Holmes 1998). We assume there is no measurement error in X (which could lead to false matches). In this case, there will be F_j individuals in the population that match the specified record. Assuming symmetry of the sampling scheme (see Section 6 for a more precise statement of the necessary condition), as for example for simple random sampling or

Bernoulli sampling, the probability that an observed match between this specified record and an individual in the population is correct, conditional on $X = j$ and F_j, is

$$\Pr(\text{correct match} \,|\, \text{unique match}, \; X = j, F_j) = 1/F_j$$

In practice, F_j will generally be unknown. We therefore consider specifying a model which generates the F_j, $j = 1, \ldots, J$, and define the record level measure of risk for a specified sample unique record with $X = j$ as

$$\theta_j = \Pr(\text{correct match} \,|\, \text{unique match}, \; X = j) = E(1/F_j \,|\, f_j = 1) \tag{1}$$

This expectation is with respect to both the model generating the F_j and the sampling scheme. This is the same measure considered by Franconi and Polettini (2004).

The risk measure in (1) may be generalised to $E(1/F_j \,|\, f_j)$ for any record in the microdata with $f_j \geq 1$. The measure assumes that the intruder does not have "response knowledge" (Bethlehem et al. 1990), i.e., does not know whether the population individual matched to a record is in the sample or not. A conservative approach to risk assessment might seek to protect against the possibility that the intruder might collude with a respondent (or more than one respondent) to identify other respondents. If the colluding respondent could identify his or her own record in the microdata using the full set of responses he or she provided to the survey agency, the intruder could effectively remove this record from the microdata and reduce both f_j and F_j for the colluding respondent's key value j by one. Thus, if there were just one other record in the microdata with this key value then this record could effectively be treated as sample unique and its risk might therefore be judged to be increased. The possibility being considered here appears very remote, however, and we shall not pursue it any further in this article.

To implement the definition of θ_j in (1) in practice, we need to specify the model generating the F_j. Following Bethlehem et al. (1990) and other authors, we assume that the F_j are independently Poisson distributed with means λ_j treated initially as fixed parameters. We assume further, like Skinner and Holmes (1998), that the sampling scheme is such that f_j and $z_j = F_j - f_j$ are independently Poisson distributed as

$$f_j \,|\, \lambda_j \sim Po(\pi\lambda_j) \quad \text{and} \quad z_j \,|\, \lambda_j \sim Po((1-\pi)\lambda_j) \tag{2}$$

This is the case, for example, under Bernoulli sampling with selection probability π. We let $\mu_j = \pi\lambda_j$ and note that the μ_j represent the expected sample frequencies in the m-way contingency table formed by cross-classifying the key variables $X_1, \ldots, X_m$. There are different possible approaches to modelling the μ_j (or equivalently λ_j). One approach is to treat them as independent and identically distributed (*iid*) outcomes of a random variable, as in Bethlehem et al. (1990). This approach implies, however, that the μ_k, for $k \neq j$, will carry no information about the risk measure θ_j so that the intruder will be unable to use the information provided by the f_k for other key values (i.e., $k \neq j$) to judge whether θ_j should be relatively high or low. This approach may therefore fail to identify particularly risky records. We therefore consider two broader classes of models, which include the *iid* model as a special case. The first class of models consists of log-linear models of a conventional form (e.g., Agresti 1996) in which the μ_j are generated by

$$\log \mu_j = x_j{}'\beta \tag{3}$$

where x_j is a vector containing specified main effects and interactions for $X_1, \ldots, X_m$. Note that, under this model, the expected population frequencies, λ_j, also follow a log-linear model differing from (3) only by an intercept term (which is decreased by $\log \pi$).

This model assumes certain relationships between the μ_j for different j and hence enables inference about the risk measure θ_j for a given key value j to borrow information provided by the f_k for other key values. For example, under an independence model with just two key variables, X_1 and X_2, the smallest value of μ_j will correspond to the rarest categories for these two variables, as implied by the relevant elements of β. Evidence about the relative size of μ_j (and therefore of θ_j) in this example will be provided not only by f_j but also by the f_k for key values k for which the category of either X_1 or X_2 corresponds to the categories defining j (these values are used in the estimation of β). If the categories of both X_1 and X_2 are relatively rare and if $f_j = 1$, we may expect θ_j to be relatively high.

A key issue in our approach is the choice of which terms x_j to include in (3) and, in particular, how ‘complex’ a model to specify. If the model includes many higher order interactions, then, as we shall note in Section 3.2., the resulting estimated measures of risk may be either unstable or not very informative. On the other hand, if we over-simplify the model by omitting important interaction terms the estimated risk measures may fail to capture all the variation between the μ_j and hence fail to identify records which are particularly risky. We suggest that an appropriate theoretical criterion for model selection in disclosure risk assessment is that it provides ‘good’ prediction of the $1/F_j$ in (1). This is not the same as the more conventional criterion of good fit to the data, although the two criteria are likely to be related in practice. How to translate the theoretical criterion of good prediction into a data-based criterion is beyond the scope of this article, however. Here, we shall simply address the model choice issue empirically, firstly by considering two alternative specifications of the x_j as discussed in Section 5 and, secondly, by making allowance for the possibility that a given specification of (3) fails to capture all the variation between the μ_j by generalising the model to include a random effect ε_j as follows:

$$\log \mu_j = x_j{}'\beta + \varepsilon_j \tag{4}$$

allowing for departures from the model in (3) via possible overdispersion. See, for example, Cameron and Trivedi (1998) and Agresti (1996). Such a model has been considered in the disclosure control context by Skinner and Holmes (1998).

For simplicity, we specify a gamma distribution for $\omega_j = \exp(\varepsilon_j)$ as

$$g(\omega; v, b) = \frac{b^v}{\Gamma(v)}\omega^{v-1}\exp(-b\omega), \quad v, b > 0$$

where $E(\omega) = v/b$ and $\mathrm{var}(\omega) = v/b^2$. To centre the distribution of ε_j, the gamma mean is assumed to be one, $\upsilon = b$, that is

$$g(\omega_j; \upsilon) = \frac{\upsilon^{\upsilon}}{\Gamma(\upsilon)} \omega_j^{\upsilon-1} \exp(-\upsilon\omega_j) \tag{5}$$

Given the model specified by (2) and either (3) or (4), we return to consideration of the record level measure of risk, defined in (1). We may write

$$\theta_j = E\left[\frac{1}{f_j + z_j} | f_j = 1, \, data\right] = E\left[E\left(\frac{1}{f_j + z_j} | \lambda_j\right) | f_j = 1, \, data\right] \tag{6}$$

It follows from (2) that

$$\begin{aligned} E\left(\frac{1}{f_j + z_j} | \lambda_j\right) &= \sum_{z=0}^{\infty} \frac{1}{(1+z)} \frac{\exp[-(1-\pi)\lambda_j][(1-\pi)\lambda_j]^z}{z!} \\ &= \frac{1}{(1-\pi)\lambda_j}\{1 - \exp[-(1-\pi)\lambda_j]\} \end{aligned} \tag{7}$$

Under Model (3), λ_j is fixed and so (7) provides an expression for θ_j. Under Model (4), λ_j is random and we obtain from (6) and (7) that

$$\begin{aligned} \theta_j &= E\left[\frac{1}{(1-\pi)\lambda_j}\{1 - \exp[-(1-\pi)\lambda_j]\} | f_j = 1, \, data\right] \\ &= \int \frac{1}{(1-\pi)\lambda_j}\{1 - \exp[-(1-\pi)\lambda_j]\} g(\lambda_j | f_j = 1) d\lambda_j \end{aligned} \tag{8}$$

where $g(\lambda_j | f_j = 1)$ is the conditional density of λ_j given that $f_j = 1$. Under the gamma model in (5), we may write

$$\theta_j = \int_0^{\infty} \frac{1}{(1-\pi)\pi^{-1}\omega\phi_j}\{1 - \exp[-(1-\pi)\pi^{-1}\omega\phi_j]\} g(\omega | f_j = 1) d\omega \tag{9}$$

where $\phi_j = \exp(x_j{}'\beta)$. From Skinner and Holmes (1998) we find that

$$g(\omega_j | f_j = 1) = \frac{\mu_j \exp(-\mu_j) g(\omega_j)}{\int \mu_j \exp(-\mu_j) g(\omega_j) d\omega_j} \tag{10}$$

and for the gamma model we find that the conditional distribution of ω_j given $f_j = 1$ is also gamma with parameters $\upsilon + 1$ and $\upsilon + \phi_j$. It follows from (9) and (10) that

$$\theta_j = \frac{\pi(\phi_j + \upsilon)}{(1-\pi)\phi_j\upsilon}\left[1 - \left(\frac{\phi_j + \upsilon}{\pi^{-1}\phi_j + \upsilon}\right)^{\upsilon}\right] \tag{11}$$

3.2. Estimation of θ_j Under Log-Linear Model in (3)

We assume that the F_j are unobserved and that the data available to estimate θ_j consist of the sample frequencies f_j. From (2) these are assumed to be independently Poisson distributed, $f_j \sim P_O(\mu_j)$ with the μ_j obeying (3). The parameter vector β in (3) may be estimated by maximum likelihood using iterative proportional fitting (Agresti 1996), to give an estimated vector $\hat{\beta}$ and fitted values $\hat{\mu}_j = \exp(x_j{}'\hat{\beta})$. From (6) and (7) the estimated disclosure risk is

$$\hat{\theta}_j = \frac{1}{(1-\pi)\hat{\lambda}_j}\{1 - \exp[-(1-\pi)\hat{\lambda}_j]\}$$

$$= \frac{1}{(1-\pi)\pi^{-1}\hat{\mu}_j}\{1 - \exp[-(1-\pi)\pi^{-1}\hat{\mu}_j]\} \qquad (12)$$

As mentioned in Section 3.1., the $\hat{\mu}_j$ and hence the $\hat{\theta}_j$ may be unstable or not very informative if many higher order interaction terms are included in (3). In the extreme case, if a saturated model is employed, $\hat{\mu}_j = 1$ for all j and the $\hat{\theta}_j$ fail to discriminate at all between the sample unique cases. On the other hand, a model with too few interaction terms may fail to capture the full variation between the μ_j. To allow for this we now consider estimation under the overdispersed Model (4).

3.3. *Estimation of θ_j Under Overdispersed Log-Linear Model in (4)*

We again estimate the parameters β and v of the model defined by (4) and (5) by maximum likelihood to give $\hat{\beta}$ and $\hat{v}$. We let $\hat{\phi}_j = \exp(x_j{}'\hat{\beta})$ and plugging $\hat{\phi}_j$ and $\hat{v}$ into the expression in (11) we obtain, as our estimated risk measure

$$\hat{\theta}_j = \frac{\pi(\hat{\phi}_j + \hat{v})}{(1-\pi)\hat{\phi}_j\hat{v}}\left[1 - \left(\frac{\hat{\phi}_j + \hat{v}}{\pi^{-1}\hat{\phi}_j + \hat{v}}\right)^{\hat{v}}\right]$$

4. Relation to File Level Measures of Risk

Four file level measures of risk considered in the literature are:

$$\Pr(PU) = \sum I(f_j = 1, F_j = 1)/n$$

$$\Pr(PU|SU) = \sum I(f_j = 1, F_j = 1)/\sum I(f_j = 1)$$

$$\theta_U = \sum I(f_j = 1)/\sum F_j I(f_j = 1) \quad \text{and}$$

$$\theta_s = \sum F_j^{-1} I(f_j = 1)/\sum I(f_j = 1)$$

where all the summations are over $j = 1, \ldots, J$. The first two measures may be interpreted as the proportions of sample individuals or sample unique individuals, respectively, that are population unique (Fienberg and Makov 1998; Samuels 1998). Since only sample unique records can be population unique we must have $\Pr(PU) \leq \Pr(PU|SU)$ and the latter measure may be treated as more conservative. Skinner and Elliot (2002) argue, however, that both these measures may be overoptimistic, because they fail to reflect the risk arising from values of X which are twins ($F_j = 2$), triples ($F_j = 3$) and so forth, and they introduce the third and fourth measures. These may be interpreted as the probability that an observed match (on the key variables) between a sample unique individual and a known individual in the population is in fact correct, according to whether the individual is drawn at random (with equal probability) from the population, for θ_U, or from the sample unique cases, for θ_s. Whether θ_U or θ_s is a more realistic measure depends upon the assumed threat from the intruder, but it will always be the case that $\theta_U \leq \theta_s$.

The file level measures above may all be interpreted as probabilities with respect to sampling mechanisms which draw individuals from the population or sample with equal probability. These probabilities are effectively unconditional on the value of X. In contrast, the record level measure in Section 3.1 may be interpreted as a probability conditional on the values of the key variables defining X.

The measure θ_j has the same form as the file level measures θ_U and θ_s if the expectation in (1) is replaced by a mean of F_j^{-1} across sample unique records, either with weights proportional to F_j for θ_U or with equal weights for θ_s. In particular, we may expect that the (unweighted) average of the record level measures θ_j will approximately equal θ_s. Since $\theta_s \geq \theta_U$, it follows that if θ_U is used as a file level measure, e.g., for the reasons of simplicity of estimation discussed in Skinner and Elliot (2002), this measure will tend to understate the (unweighted) average of the record level measures of risk θ_j.

5. Empirical Evaluation

In this section we seek to evaluate the properties of the $\hat{\theta}_j$ empirically using an artificial finite population. We wish to avoid basing our evaluation on any single assumed model and hence cannot simply compare the values of $\hat{\theta}_j$ with "true values" θ_j, since the latter are defined with respect to a model. We therefore adopt two alternative approaches. First, we study the relation between $\hat{\theta}_j$ and the empirical proportion of population uniques among sample unique units. Second, we study the relation between the average value of $\hat{\theta}_j$ and the average value of $1/F_j$ overall and within subgroups. For $\hat{\theta}_j$ to be a useful measure, we expect a strong positive relationship in the first case and a strong positive relationship, with approximate equality between the two averages, in the second case.

As a basis for studying these relationships, we constructed an artificial population file by combining data for two years (1996, 1997) from the UK General Household Survey, resulting in records on $N = 33,142$ individuals. Following consideration of possible intruder scenarios by Dale and Elliot (2001), we used the following $m = 5$ key variables:

1. X_1, sex in 2 categories
2. X_2, marital status in 7 categories
3. X_3, economic status in 13 categories
4. X_4, socio-economic group in 10 categories
5. X_5, age in ten-year bands in 8 categories

generating $J = 2 \times 7 \times 13 \times 10 \times 8 = 14,560$ possible key values. We evaluated the estimated measures of disclosure risk for two simple random samples from this population, one of size $n = 2,500$ ($\pi = 0.075$) and one of size $n = 5,000$ ($\pi = 0.15$). The numbers of sample uniques were $n_1 = 370$ in the first sample and $n_1 = 495$ in the second sample. The corresponding numbers of population uniques in these samples were 59 and 130 respectively. The four file level measures of risk (see, Section 4) were:

- sample 1 $(n = 2,500)$: $\Pr(PU) = 0.024$, $\Pr(PU \mid SU) = 0.159$, $\theta_U = 0.115$, $\theta_s = 0.313$;
- sample 2 $(n = 5,000)$: $\Pr(PU) = 0.026$, $\Pr(PU \mid SU) = 0.262$, $\theta_U = 0.210$, $\theta_s = 0.443$.

As expected, we find $\Pr(PU) \leq \Pr(PU \mid SU) \leq \theta_s$ and $\theta_U \leq \theta_s$ for both samples so that θ_s is the most conservative measure.

We next compute values of $\hat{\theta}_j$ for each of the sample unique cases in each sample. We first consider the log-linear model in (3) for which the λ_j are fixed and obtain the $\hat{\theta}_j$ as discussed in Section 3.2. We consider the following two specifications of the model in (3):

- Model 1: a log-linear model including all main effects;
- Model 2: a log-linear model including also all two-factor interactions.

Two specifications are considered to allow some empirical assessment of the effect of model choice. The independence model, Model 1, is chosen as a simple model which can be easily fitted in practice and which allows the risk measure to reflect the relative rarity of the categories in each individual key variable separately, if not in combination. Model 2 allows the risk measure to capture unusual combinations of categories of pairs of key variables. It is more demanding computationally to estimate than Model 1 but is still sufficiently parsimonious for overfitting not to appear to be a problem. The only zero counts observed in the cells of two-way margins of the table cross-classifying the key variables are structural zeros. On the other hand there are many sampling zeros in the three-way margins of this table, and this suggests that including three-factor interactions in the model might lead to overfitting. Model choice is discussed further in Sections 3.1 and 6.

Tables 1–4 show the distributions of $\hat{\theta}_j$ across sample unique cases for these two models for both samples. For the first sample $(n = 2,500)$, we find the mean values of $\hat{\theta}_j$ to be 0.442 and 0.296 for Models 1 and 2 respectively, compared with the "expected" mean $\theta_s = 0.313$. For the second sample $(n = 5,000)$ we find mean values of $\hat{\theta}_j$ of 0.513 and 0.435 for the two models, compared with $\theta_s = 0.443$. The correspondence with θ_s seems rather better for Model 2. (This suggests a means of estimating θ_s to augment the simpler approach to estimating θ_U discussed by Skinner and Elliot (2002).) In all cases θ_U understates substantially the average record level measure.

The five divisions of the range $[0, 1]$ for $\hat{\theta}_j$ in Tables 1 and 2 define subsets of sample uniques with similar values of $\hat{\theta}_j$. For each of these subsets, the proportion of population unique cases is presented in these tables. As in Skinner and Holmes (1998), we find that $\hat{\theta}_j$ are useful for deciding whether a sample unique case is population unique, with Model 2 providing better discrimination. For the first sample, it is more likely than not that a sample unique is population unique if $\hat{\theta}_j > 0.8$ for Model 2, but not for Model 1. The ability to detect population uniques with high probability is even stronger for the second sample.

Tables 3 and 4 give the results when λ_j is random and follows a gamma distribution, as discussed in Section 3.3. We find similar results to the model with no overdispersion, with no evidence of improved discrimination for the model with random effects.

We next consider the relation between the average value of $1/F_j$ and $\hat{\theta}_j$. Table 5 presents the means of $1/F_j$ within each of the five classes of values of $\hat{\theta}_j$ considered in the

Table 1. Frequencies and proportions of population unique cases for sample unique records within classes of values of $\hat{\theta}_j$ for Models 1 and 2 with no overdispersion and $n = 2,500$

Range of values of $\hat{\theta}_j$	Model 1		Model 2	
	Frequency	Proportion population unique	Frequency	Proportion population unique
(0.0, 0.2)	84	0.07	113	0.07
(0.2, 0.4)	61	0.11	68	0.08
(0.4, 0.6)	88	0.13	78	0.09
(0.6, 0.8)	79	0.19	67	0.18
(0.8, 1.0)	58	0.33	44	0.59
Total	370		370	

previous tables. Given the lack of evidence of improved performance using random effects, we only consider the model with λ_j fixed. For the interpretation of $\hat{\theta}_j$ to be valid, we expect the mean values to lie approximately within the class intervals. As before, we find that the results for Model 2 show greater validity, with all the average values of $1/F_j$ falling within the class intervals of $\hat{\theta}_j$.

Table 5 provides empirical verification of the interpretation of $\hat{\theta}_j$ as the probability of a correct match. To check this validity further, we have also studied the relationship between the mean of $\hat{\theta}_j$ and the mean of $1/F_j$ within the $40\,(= 2 + 7 + 13 + 10 + 8)$ subgroups defined by the univariate categories of the five key variables for sample unique records for each of the two samples. Table 6 gives the results for the seven subgroups defined by the categories of marital status for Model 2. We observe a good correspondence between the two sets of means, especially for the larger sample size. For the purpose of disclosure control we might wish to use the values of $\hat{\theta}_j$ to identify any categories with a high level of risk. It is clear from the $1/F_j$ columns that Category 7 is the riskiest, and calculation of the mean of the $\hat{\theta}_j$ among sample uniques within this category would enable the risky nature of this category to be identified. In fact, for the case $n= 2,500$, there are just two sample uniques in this category, both population unique, and this high level of risk is captured by the average value of 0.93 of $\hat{\theta}_j$ for these two records. For the case $n = 5,000$, there are six sample uniques in this category. The probability that a match

Table 2. Frequencies and proportions of population unique cases for sample unique records within classes of values of $\hat{\theta}_j$ for Models 1 and 2 with overdispersion and $n = 2,500$

Range of values of $\hat{\theta}_j$	Model 1		Model 2	
	Frequency	Proportion population unique	Frequency	Proportion population unique
(0.0, 0.2)	79	0.05	105	0.06
(0.2, 0.4)	64	0.08	86	0.06
(0.4, 0.6)	85	0.15	79	0.10
(0.6, 0.8)	87	0.22	59	0.27
(0.8, 1.0)	55	0.34	41	0.58
Total	370		370	

Table 3. Frequencies and proportions of population unique cases for sample unique records within classes of values of $\hat{\theta}_j$ for Models 1 and 2 with no overdispersion and $n = 5,000$

Range of values of $\hat{\theta}_j$	Model 1		Model 2	
	Frequency	Proportion population unique	Frequency	Proportion population unique
(0.0, 0.2)	110	0.11	137	0.07
(0.2, 0.4)	94	0.11	92	0.08
(0.4, 0.6)	98	0.12	88	0.14
(0.6, 0.8)	92	0.42	76	0.49
(0.8, 1.0)	101	0.55	92	0.70
Total	495		495	

between one of these records and a population individual would be correct is 0.88, and this high risk is again captured well by the mean value of $\hat{\theta}_j$ of 0.82. Considering all 40 subgroups we found that the correlation coefficients between the two means are 0.76 and 0.82 for Models 1 and 2 respectively, with $n = 2,500$ and 0.75 and 0.96 for the models with $n = 5,000$. It is again clearly preferable to include the two-way interactions in the model.

Figure 1 displays scatterplots of values of $(1/F_j, \hat{\theta}_j)$ for the 40 subgroups with 45-degree lines joining (0, 0) and (1, 1) superimposed. They confirm the closeness between the means of $\hat{\theta}_j$ and $1/F_j$, especially for Model 2.

6. Conclusion

Skinner and Elliot (2002) argued in favour of measuring disclosure risk at the file level by the probability that an observed match is correct rather than by the probability of population uniqueness. In this article, we have shown how the record level measure of disclosure risk of Skinner and Holmes (1998), defined in terms of the probability of population uniqueness, may be extended in a parallel way to a record level measure of the probability that an observed match is correct. Both measures depend on the specification of a log-linear model for an assumed set of key variables. In an empirical evaluation of different versions of the new record level measure using real survey data, we found

Table 4. Frequencies and proportions of population unique cases for sample unique records within classes of values of $\hat{\theta}_j$ for Models 1 and 2 with overdispersion and $n = 5,000$

Range of values of $\hat{\theta}_j$	Model 1		Model 2	
	Frequency	Proportion population unique	Frequency	Proportion population unique
(0.0, 0.2)	88	0.09	114	0.08
(0.2, 0.4)	123	0.17	146	0.20
(0.4, 0.6)	102	0.23	111	0.23
(0.6, 0.8)	99	0.32	83	0.45
(0.8, 1.0)	83	0.54	41	0.71
Total	495		495	

Table 5. Means of $1/F_j$ within classes of values of $\hat{\theta}_j$ for Models 1 and 2 with no overdispersion

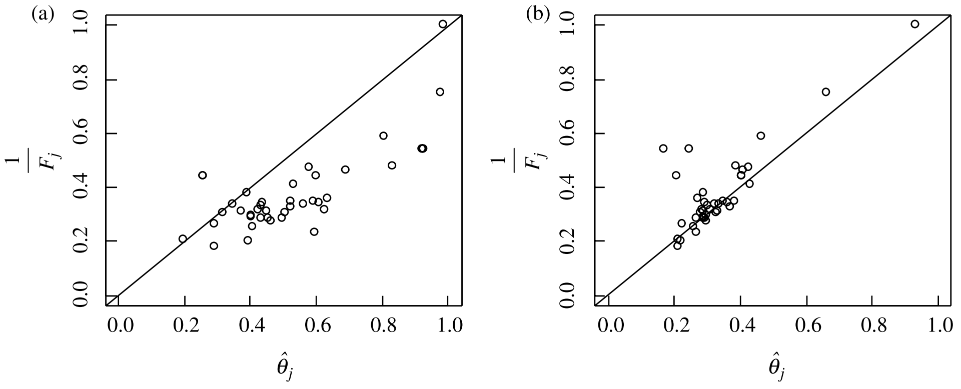
Range of values of $\hat{\theta}_j$	$n = 2,500$		$n = 5,000$	
	Model 1	Model 2	Model 1	Model 2
(0.0, 0.2)	0.06	0.09	0.07	0.06
(0.2, 0.4)	0.25	0.28	0.22	0.29
(0.4, 0.6)	0.38	0.42	0.36	0.44
(0.6, 0.8)	0.53	0.59	0.58	0.61
(0.8, 1.0)	0.68	0.81	0.73	0.83

evidence of discrimination by the measure between records of different levels of risk; in particular records which are very likely to be population unique could be identified by consideration of records with high values of the measure. We found no evidence, however, that allowance for overdispersion via the inclusion of random effects in the model improved its performance. The measure obtained under the simpler model with no random effects was validated by comparing its average value in 40 subpopulations with the "true" population quantity it was estimating, and the relationship was found to be very good for a model including only one- and two-way interactions. This measure is much easier to compute, requiring only the fitting of a standard log-linear model, than the measure proposed by Skinner and Holmes (1998), which additionally required numerical integration. In summary, we suggest for use in practice the measure obtained from Equation (12) for a log-linear model with main effects and two-way interactions. We are currently exploring the robustness of the measure to model choice and whether any improvements can be obtained through the use of higher-order interactions and model selection techniques. The questions of how model selection should depend upon the sizes of the sample and the population and whether models should be fitted separately for different subpopulations also need consideration. It may be appropriate to use simpler models for smaller sample sizes.

The measure obtained from (12) ignores any error in estimating the parameters β of the log-linear model by $\hat{\beta}$. In principle, if the true measure is taken as the posterior probability of a correct match from a Bayesian perspective and if uncertainty about β can be represented in an appropriate way (this may need to take account of the complexity of the

Table 6. Means of $\hat{\theta}_j$ and $1/F_j$ across seven subsets of the sample unique records defined by the categories of marital status, for Model 2 without overdispersion

Marital status category	$n = 2,500$		$n = 5,000$	
	Mean $\hat{\theta}_j$	Mean $1/F_j$	Mean $\hat{\theta}_j$	Mean $1/F_j$
1	0.22	0.27	0.36	0.38
2	0.39	0.35	0.54	0.50
3	0.30	0.29	0.41	0.41
4	0.22	0.20	0.36	0.36
5	0.35	0.35	0.48	0.49
6	0.41	0.47	0.53	0.54
7	0.93	1.00	0.82	0.88



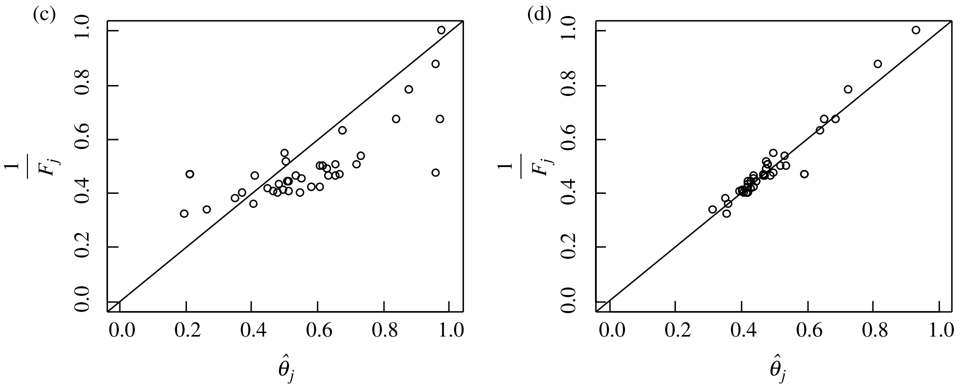


Fig. 1. Scatter plots of means of $1/F_j$ against means of estimated measure of risk $\hat{\theta}_j$ with $y = x$ lines for (a) Model 1 with $n = 2,500$, (b) Model 2 with $n = 2,500$, (c) Model 1 with $n = 5,000$, (d) Model 2 with $n = 5,000$

survey sampling scheme) then this uncertainty could be integrated out, perhaps using a simulation-based approach. We have not pursued this possibility, however, and suspect that it is more important initially to explore the dependence of the measure on model specification.

This article has assumed that the key variables are categorical and that the intruder measures X in the same way it is recorded in the microdata. In separate work, we are considering the effect of measurement error arising from the application of a known misclassification matrix to X in the microdata. This article has some conceptual relevance to the case of continuous key variables in the sense that the definition of risk as the expected value of $1/F_j$ applies for general key variables if it is assumed that the intruder employs a matching algorithm for which F_j is the number of individuals in the population that the intruder would match to a given microdata record (and it is assumed that these individuals include the true respondent). However, the log-linear modelling approach in this article only relates to the case of categorical key variables.

This article has also made simplifying assumptions about the sampling scheme. A number of issues arise in considering the possible effect of complex sampling. The first is whether the measure in (1) should be modified since the assumption that $1/F_j$ is the probability of a correct match given a unique match (see Section 3.1) depends upon assumptions about the sampling scheme. Thus, suppose that the F_j population units with key value j are ordered $k = 1, \ldots, F_j$ and suppose that the intruder seeks a microdata record which matches the kth of these units. Let A_k be the event that the kth unit is sampled and the remaining $F_j - 1$ units are not sampled. Then the probability of a correct match given a unique match and known F_j is the probability of A_k given that one of $A_1, \ldots, A_{F_j}$ occurs. In general this will depend upon F_j-way inclusion probabilities and thus be complex. In many practical circumstances, however, it might be natural to include the design variables, according to which inclusion probabilities vary amongst the key variables, since these design variables are often visible. In this case, it may be reasonable to assume that $A_1, \ldots, A_{F_j}$ are equiprobable and thus that the probability $1/F_j$ still applies. This would be similar to the approach of Franconi and Polettini (2004), who assume that all units with a common key value have a common inclusion probability and that the definition in (1) is still appropriate. On the other hand, if inclusion probabilities do vary among units with common values of the key variables and if these probabilities could be determined by the intruder for population units and sample design weights are included in the microdata file, then the identification risk may be seriously increased (de Waal and Willenborg 1997).

Even if the measure in (1) is retained, there is still the question of how the complex design should be allowed for in Expressions (7) and (8) and in the estimation method. As a pragmatic approach we suggest retaining Expressions (7) and (8), with π now being the actual inclusion probability for the record, which may vary between records. Regarding estimation, since the aim is to predict the finite population quantities, F_j, we suggest that a log-linear model for the λ_j be fitted using standard survey weighting (Rao and Thomas 2003). Complex design issues require further research, however. See Skinner and Carter (2003) on the use of weights in the file level measure of Skinner and Elliot (2002) and Franconi and Polettini (2004) on the use of auxiliary population information via calibration weights in record level measures.

7. References

Agresti, A. (1996). An Introduction to Categorical Data Analysis. New York: Wiley.

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

Cameron, C.A. and Trivedi, P.K. (1998). Regression Analysis of Count Data. Cambridge.

Carlson, M. (2002). Assessing Microdata Disclosure Risk Using the Poisson-inverse Gaussian Distribution. Statistics in Transition, 901–925.

Dale, A. and Elliot, M. (2001). Proposals for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk. Journal of the Royal Statistical Society, Series A, 164, 427–447.

de Waal, A.G. and Willenborg, L.C.R.J. (1997). Statistical Disclosure Control and Sampling Weights. Journal of Official Statistics, 13, 417–434.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (eds) (2001). Confidentiality Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland.

Duncan, G. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207–217.

Elliot, M. (2001). Disclosure Risk Assessment. In Confidentiality Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds). North-Holland, 75–90.

Fienberg, S. and Makov, U. (1998). Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data. Journal of Official Statistics, 14, 385–397.

Franconi, L. and Polettini, S. (2004). Individual Risk Estimation in (-ARGUS: a Review. In J. Domingo-Ferrer and V. Torra (eds). Privacy in Statistical Databases, Springer Lecture Notes in Computer Science 3050. Berlin, 262–272.

Fuller, W.A. (1993). Masking Procedures for Microdata Disclosure Limitation. Journal of Official Statistics, 9, 383–406.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–331.

Paass, G. (1987). Disclosure Risk and Disclosure Avoidance for Microdata. Journal of Business and Economic Statistics, 6, 487–500.

Rao, J.N.K. and Thomas, D.R. (2003). Analysis of Categorical Response Data from Complex Surveys: an Appraisal and Update. In Analysis of Survey Data, R.L. Chambers and C.J. Skinner (eds) Chichester: Wiley, 85–108.

Samuels, S. (1998). A Bayesian, Species-Sampling-Inspired Approach to the Uniques Problems in Microdata Disclosure Risk Assessment. Journal of Official Statistics, 14, 373–383.

Skinner, C.J. and Carter, R.G. (2003). Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling. Survey Methodology, 29, 177–180.

Skinner, C.J. and Elliot, M. (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. New York: Springer.

Received July 2004
Revised September 2005