Record-level Measures of Disclosure Risk for Survey Microdata

Elsayed A. H. Elamir and Chris J. Skinner

Abstract

Measures of disclosure risk at the record level have a variety of potential uses in statistical disclosure control for microdata. We propose a new 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 model and a Poisson-gamma model. Moreover, we apply the approaches to a sample of microdata from the U.K. General Household Survey. The results indicate that the risk measure may be used to establish whether sample unique records are unique in the population.

S^{3}RI Methodology Working Paper M04/02

Record-level Measures of Disclosure Risk for Survey Microdata

Elsayed A. H. Elamir and Chris J. Skinner
cjs@socsci.soton.ac.uk and eahe@socsci.soton.ac.uk

Southampton Statistical Sciences Research Institute,
University of Southampton, Southampton,
SO17 1BJ, U.K.

Abstract

Measures of disclosure risk at the record level have a variety of potential uses in statistical disclosure control for microdata. We propose a new 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 model and a Poisson-gamma model. Moreover, we apply the approaches to a sample of microdata from the U.K. General Household Survey. The results indicate that the risk measure may be used to establish whether sample unique records are unique in the population.

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 Waal (2001) and there is increasing interest in applying these methods, especially in government statistical agencies; see Doyle et al. (2001).

In this paper 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 requiring 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* or *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 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 which have a unique combination of values of the key variables (assumed categorical) in the population; see, 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 $10,000$ 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; see, 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; see, 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.

Skinner and Holmes (1998) propose one approach to the estimation of record-level measures. 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. They define their measure as the probability of population uniqueness, with probability interpreted with respect to a model. 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.

In this paper we investigate an alternative approach. We propose a different measure, replacing the probability of population uniqueness by the probability that an observed match between a microdata record and an identifiable unit in the population is correct. This parallels the approach to file-level measures developed by Skinner and Elliot (2002) and we discuss this further in Section 2, where we also introduce the formal framework for this paper. The estimation of this new measure is discussed in Section 3, with particular consideration of how the computations in Skinner and Holmes (1998) can be simplified. An empirical evaluation of the approaches outlined in Section 3 is presented in Section 4 based upon data from the U.K. General Household Survey.

2 Framework and Disclosure Risk at the File Level

In this section we introduce the formal framework and some file-level measures of disclosure risk. 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 $s \subseteq U$ 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 $X_1, X_2, ..., X_m$.

In order to define some measures of disclosure risk we introduce some further notation. Let the variable formed by cross-classifying $X_1, X_2, ..., X_m$ be denoted X, with values denoted $1, ..., J$, 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. 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 $j = 1, ..., J$, and let the population frequencies of frequencies be $N_r = \sum_{j=1}^{J} I\left(F_j = r\right)$, $r = 1, 2, ...$ 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 $n_r = \sum_{j=1}^{J} \mathrm{I}\left(f_j = r\right)$, $r = 1, 2, ...$ For example, n_1 is the number of sample uniques.

Four examples of file level measures of risk are

$$\Pr\left(\mathrm{PU}\right)=\sum \mathrm{I}(f_j=1,F_j=1)/n,$$

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

$$\theta_U=\sum \mathrm{I}\left(f_j=1\right)\Big/\sum F_j\mathrm{I}\left(f_j=1\right),$$

and

$$\theta_s=\sum F_j^{-1}\mathrm{I}\left(f_j=1\right)\Big/\sum \mathrm{I}\left(f_j=1\right),$$

where all the summations are over $j=1,.....,J$. The first two measures may be interpreted as the proportions of sample individuals or sample unique individuals, respectively, which are population unique; see, for example, Fienberg and Makov (1998) and 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 $\theta_U\leq\theta_s$.

3 Disclosure Risk at the Record Level

In order to define record-level measures of disclosure risk we make use of the X information available for each record. The file level measures could 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. To obtain record-level measures we propose to condition these probabilities on the values of the key variables defining X. This implies that any two records with the same value of X will have the same measure of disclosure risk. In fact, we shall only consider sample records and restrict attention to records which are sample unique, since these may be expected to be the most risky following Skinner and Holmes (1998), so that all records of interest will have different values of X .

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 which match a specified record with $X = j$. Assuming symmetry of the sampling scheme, 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

$$\Pr\left(\text{correct match}|\text{ unique match, } X = j, F_j\right) = \frac{1}{F_j}.$$

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

$$\begin{aligned} \theta_j &= \Pr\left(\text{correct match}|\text{ unique match}, X = j\right) \\ &= \mathrm{E}\left(\frac{1}{F_j}\middle| f_j = 1\right) \qquad (1) \end{aligned}$$

This expectation is with respect to both the model generating the F_j and the sampling scheme.

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.

To implement the definition of θ_j 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

$$f_j \mid \lambda_j \sim \mathrm{Po}\left(\pi\lambda_j\right) \text{ and } z_j \mid \lambda_j \sim \mathrm{Po}\left[\left(1-\pi\right)\lambda_j\right]. \tag{2}$$

This is the case, for example, under Bernoulli sampling with selection probability π. It follows that

$$\begin{aligned} \theta_j &= \mathrm{E}\left[\frac{1}{f_j + z_j} \mid f_j = 1, \mathrm{data}\right] \\ &= \mathrm{E}\left[\mathrm{E}\left(\frac{1}{1+z_j} \mid \lambda_j\right) \mid f_j = 1, \mathrm{data}\right]. \end{aligned} \tag{3}$$

It follows from (2) that

$$\begin{aligned} \mathrm{E}\left(\frac{1}{1+z_j} \mid \lambda_j\right) &= \sum_{z=0}^{\infty} \frac{1}{1+z} \frac{\exp\left[-\left(1-\pi\right)\lambda_j\right]\left(\left(1-\pi\right)\lambda_j\right)^z}{z!} \\ &= \frac{1}{\left(1-\pi\right)\lambda_j}\left\{1 - \exp\left[-\left(1-\pi\right)\lambda_j\right]\right\}. \end{aligned} \tag{4}$$

If λ_j is fixed then (4) provide an expression for θ_j. If λ_j is random, we obtain from (3) and (4) that

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

where $g\left(\lambda_j \middle| f_j = 1\right)$ is the conditional density of λ_j given that $f_j = 1$. We now consider the estimation of θ_j from sample data.

3.1 Estimation of θ_j - Fixed λ_j

We assume that the F_j are unobserved and that the data available to to estimate θ_j consist of the sample frequencies f_j. From (2), these are assumed to be independently Poisson distributed, $f_j \sim \text{Po}\,(\mu_j)$, where $\mu_j = \pi\lambda_j$. A log-linear model for the μ_j may be expressed as

$$\log\mu_j = u_j'\beta \tag{5}$$

where u_j is a vector containing specified main effects and interactions for $X_1, ..., X_m$. Such a model may be fitted using standard procedures; see, Agresti (1996), to give an estimated vector $\widehat{\beta}$ and fitted values

$$\widehat{\mu}_j = \exp\left(u_j'\widehat{\beta}\right).$$

From (4) the estimated disclosure risk is

$$\begin{aligned}\widehat{\theta}_j &= \frac{1}{(1-\pi)\,\widehat{\lambda}_j}\left\{1-\exp\left[-(1-\pi)\,\widehat{\lambda}_j\right]\right\} \\ &= \frac{1}{(1-\pi)\,\pi^{-1}\widehat{\mu}_j}\left\{1-\exp\left[-(1-\pi)\,\pi^{-1}\widehat{\mu}_j\right]\right\} \qquad (6)\end{aligned}$$

If a very complex log-linear model is chosen then the resulting $\widehat{\theta}_j$ may either be unstable or not very informative. In the extreme case, if a saturated model is employed, $\widehat{\mu}_j = 1$ for all j and the $\widehat{\theta}_j$ fail to discriminate at all between the sample unique cases. This suggests selecting a simpler log-linear model. The problem then is that, if the model is 'too' simple, the specified u_j may fail to capture all the variation between the μ_j, that is there may be overdispersion. Making allowance for overdispersion in $\widehat{\theta}_j$ is discussed in the next section.

3.2 Estimation of θ_j - Random λ_j

A common approach to allowing for overdispersion is by introducing a multiplicative error term; see, for example, Cameron and Trivedi (1998) and Agresti (1996). Suppose the distribution of a random count

$y = f_j$ is conditionally Poisson, that is

$$y \mid \mu_j \sim \mathrm{Po}\left(\mu_j\right),$$

where now

$$\begin{aligned} \log \mu_j &= x_j' \beta + \varepsilon_j \\ \mu_j &= \exp\left(x_j' \beta + \varepsilon_j\right) \end{aligned}$$

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

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

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

$$g\left(w_j \mid v\right) = \frac{v^v}{\Gamma\left(v\right)} w_j^{v-1} \exp\left(-v w_j\right). \tag{7}$$

The measure of disclosure risk θ_j is then given by

$$\theta_j = \int_0^\infty \frac{1}{\left(1-\pi\right)\pi^{-1} w \phi_j} \left\{1 - \exp\left[-\left(1-\pi\right)\pi^{-1} w \phi_j\right]\right\} g\left(w \mid f_j = 1\right) dw, \tag{8}$$

where $\phi_j = \exp\left(u_j' \beta\right)$.

From Skinner and Holmes (1998) we find that

$$g\left(w_j \mid f_j = 1\right) = \frac{\mu_j \exp\left(-\mu_j\right) g\left(w_j\right)}{\int \mu_j \exp\left(-\mu_j\right) g\left(w_j\right) dw_j}. \tag{9}$$

Under the gamma model given in (7), we find that the conditional distribution of w_j give $f_j = 1$ is also gamma with parameters $v + 1$ and $v + \phi_j$. It follows from (8) and (9) that

$$\theta_j = \frac{\pi\left(\phi_j + v\right)}{\left(1-\pi\right)\phi_j v}\left[1 - \left(\frac{\phi_j + v}{\pi^{-1}\phi_j + v}\right)^v\right].$$

Suppose now that the Poisson-gamma (negative binomial) model is fitted to the f_j giving estimates $\widehat{v}$ and $\widehat{\beta}$ of the parameters. Let $\widehat{\mu}_j = \widehat{\phi}_j = \exp\left(u_j'\widehat{\beta}_j\right)$ then the estimated value of θ_j is given by

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

4 Empirical Evaluation

In this section we seek to evaluate the properties of the $\widehat{\theta}_j$ 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 $\widehat{\theta}_j$ 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 $\widehat{\theta}_j$ and the empirical proportion of population uniques among sample unique units. Second, we study the relation between the average value of $\widehat{\theta}_j$ and the average value of $1/F_j$ within subgroups. For $\widehat{\theta}_j$ 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 U.K. General Household Survey, resulting in records on $N = 33142$ 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 10 categories;

5. X_5 age in ten-year bands in 8 categories;

generating $J = 2 \times 7 \times 13 \times 10 \times 8 = 14560$ possible key values. We evaluated the estimated measures of disclosure risk for two simple random samples from this population, one of size $n = 2500$ ($\pi = 0.075$) and one of size $n = 5000$ ($\pi = 0.15$).

The numbers of sample uniques were $n_1 = 370$ in the first sample and $n_1 = 495$ in the second sample. The four file-level measures of risk were:

- sample 1 ($n = 2500$) : $\Pr(PU) = 0.024, \Pr(PU|SU) = 0.159, \theta_U = 0.115, \theta_s = 0.313$;
- sample 2 ($n = 5000$) : $\Pr(PU) = 0.026, \Pr(PU|SU) = 0.262, \theta_U = 0.210, \theta_s = 0.443$.

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

We next compute values of $\widehat{\theta_j}$ for each of the sample unique cases in each sample. We first assume fixed λ_j and compute $\widehat{\theta_j}$ using iterative proportional fitting, for the following two specifications of the model in (5):

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

Tables 1, 2, 3 and 4 show the distributions of $\widehat{\theta_j}$ across sample unique cases for these two models for both samples. For the first sample ($n = 2500$), we find the mean values of $\widehat{\theta_j}$ 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 ($n = 5000$) we find mean values of $\widehat{\theta_j}$ 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.

	Model 1		Model 2	
$\widehat{\theta}_j$	Freq.	Prop. Pop. Unique	Freq.	Prop. Pop. Unique
0−	84	0.07	113	0.07
0.20−	61	0.11	68	0.08
0.40−	88	0.13	78	0.09
0.60−	79	0.19	67	0.18
0.80 − 1	58	0.33	44	0.59
Total	370		370	

Table 1: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with no overdispersion and $n = 2500$.

	Model 1		Model 2	
$\widehat{\theta}_j$	Freq.	Prop. Pop. Unique	Freq.	Prop. Pop. Unique
0−	79	0.05	105	0.06
0.20−	64	0.08	86	0.06
0.40−	85	0.15	79	0.10
0.60−	87	0.22	59	0.27
0.80 − 1	55	0.34	41	0.58
Total	370		370	

Table 2: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with overdispersion and $n = 2500$.

The five divisions of the range $[0, 1]$ for $\widehat{\theta}_j$ in Tables 1 and 2 define subsets of sample uniques with similar values of $\widehat{\theta}_j$. For each of these subsets, the proportion of population unique cases are presented in these tables. As in Skinner and Holmes (1998), we find that the $\widehat{\theta}_j$ 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 $\widehat{\theta}_j > 0.8$ 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 3.2. We find similar results to the model with no overdispersion, with no evidence of im-

	Model 1		Model 2	
$\widehat{\theta}_j$	Freq.	Prop. Pop. Unique	Freq.	Prop. Pop. Unique
0−	110	0.11	137	0.07
0.20−	94	0.11	92	0.08
0.40−	98	0.12	88	0.14
0.60−	92	0.42	76	0.49
$0.80 - 1$	101	0.55	92	0.70
Total	495		495	

Table 3: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with no overdispersion and $n = 5000$.

	Model 1		Model 2	
$\widehat{\theta}_j$	n_1	Prop. Pop. Unique	n_1	Prop. Pop. Unique
0−	88	0.09	114	0.08
0.20−	123	0.17	146	0.20
0.40−	102	0.23	111	0.23
0.60−	99	0.32	83	0.45
$0.80 - 1$	83	0.54	41	0.71
Total	495		495	

Table 4: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with overdispersion and $n = 5000$.

proved discrimination for the model with random effects.

We next study the relationship between the mean of $\widehat{\theta}_j$ 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. Tables 5 and 6 gives the results for the main effects and all two-way interaction models for $\pi = 0.075$ and 0.15. Given the lack of evidence of improved performance using random effects, we only consider the model with λ_j fixed. We find, as expected, a strong relationship between the mean of the $\widehat{\theta}_j$ and the mean of the values $1/F_j$. The two means are broadly similar for all the subgroups h, except for some cases where the size of the subgroup, n_h, is small. The correlation coefficients between the two means are 0.76 and 0.82 for the two models with $\pi = 0.075$ and 0.75 and 0.96

for the models with $\pi = 0.15$. It is clearly preferable to include the two-way interaction in the model.

Regression curves, obtained using the loess method (locally weighted regression scatter plot smoothing; see, Cleveland (1979) and Bowman and Azzalini (1997)) are displayed in Figure 1 for the data in Tables 5 and 6. They confirm the strong linear relationship between the mean of $\widehat{\theta}_j$ and the mean of $1/F_j$, especially for the model including two-way interactions.

5 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 paper, 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 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 forty 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

Variable	Subpop.	Samp. Unique	Mean of $\widehat{\theta}_j$		mean of
	h	n_{1h}	Model 1	Model 2	F_j^{-1}
Sex	1	202	0.431	0.300	0.333
	2	168	0.455	0.292	0.288
Marital status	3	108	0.291	0.224	0.266
	4	40	0.522	0.385	0.349
	5	94	0.401	0.296	0.294
	6	31	0.393	0.221	0.202
	7	62	0.593	0.347	0.351
	8	33	0.690	0.410	0.465
	9	2	0.988	0.932	1
Economic status	10	104	0.197	0.214	0.206
	11	7	0.926	0.245	0.541
	12	3	0.921	0.167	0.541
	13	33	0.506	0.327	0.308
	14	33	0.577	0.425	0.472
	15	34	0.610	0.362	0.345
	16	58	0.316	0.281	0.308
	17	38	0.597	0.269	0.235
	18	6	0.831	0.386	0.478
	19	14	0.806	0.467	0.587
	20	8	0.256	0.405	0.441
	21	2	0.977	0.661	0.75
	22	30	0.293	0.213	0.182
Socioeco. group	23	26	0.349	0.325	0.338
	24	40	0.388	0.291	0.380
	25	42	0.434	0.290	0.286
	26	46	0.374	0.287	0.310
	27	58	0.405	0.259	0.253
	28	73	0.496	0.267	0.285
	29	42	0.524	0.369	0.327
	30	8	0.256	0.405	0.441
	31	29	0.561	0.338	0.340
	32	6	0.602	0.208	0.444
Age	33	26	0.634	0.272	0.361
	34	28	0.627	0.283	0.315
	35	60	0.463	0.297	0.274
	36	72	0.403	0.288	0.292
	37	64	0.437	0.294	0.344
	38	40	0.426	0.311	0.315
	39	50	0.449	0.332	0.311
	40	30	0.531	0.431	0.409

Table 5: means of $\widehat{\theta}_j$ and $1/F_j$ across forty subpopulations (subpop.) defined by Sex (2), Marital status (7), Economic status (13), Socio-economic group (10) and Age (8) for sample unique records with models 1 and 2 and $n = 2500$.

Variable	Suppop.	Samp. Unique	Mean of $\widehat{\theta}_j$		Mean of
	h	n_{1h}	Model 1	Model 2	F_j^{-1}
Sex	1	231	0.510	0.448	0.442
	2	264	0.515	0.423	0.443
Marital status	3	119	0.352	0.355	0.379
	4	59	0.619	0.538	0.500
	5	123	0.468	0.405	0.408
	6	53	0.407	0.364	0.361
	7	80	0.628	0.476	0.488
	8	55	0.732	0.532	0.538
	9	6	0.958	0.817	0.875
Economic status	10	125	0.267	0.315	0.338
	11	5	0.958	0.500	0.475
	12	7	0.975	0.653	0.671
	13	55	0.583	0.420	0.421
	14	42	0.668	0.468	0.471
	15	56	0.656	0.490	0.465
	16	60	0.373	0.407	0.402
	17	65	0.551	0.440	0.452
	18	8	0.878	0.728	0.783
	19	32	0.840	0.688	0.674
	20	14	0.212	0.594	0.470
	21	1	0.976	0.933	1
	22	25	0.610	0.521	0.502
Socioeco. group	23	28	0.500	0.500	0.547
	24	54	0.410	0.440	0.461
	25	51	0.496	0.405	0.409
	26	72	0.45	0.429	0.418
	27	70	0.485	0.421	0.430
	28	89	0.550	0.416	0.399
	29	59	0.610	0.438	0.422
	30	14	0.212	0.594	0.470
	31	50	0.656	0.480	0.506
	32	8	0.675	0.641	0.629
Age	33	33	0.721	0.480	0.506
	34	55	0.636	0.470	0.465
	35	83	0.508	0.431	0.443
	36	100	0.514	0.397	0.406
	37	72	0.479	0.424	0.401
	38	68	0.505	0.476	0.517
	39	49	0.535	0.473	0.464
	40	35	0.196	0.358	0.324

Table 6: means of $\widehat{\theta}_j$ and $1/F_j$ across forty subpopulations (subpop.) defined by Sex (2), Marital status (7), Economic status (13), Socioeconomic group (10) and Age (8) for sample unique records with models 1 and 2 and $n = 5000$.

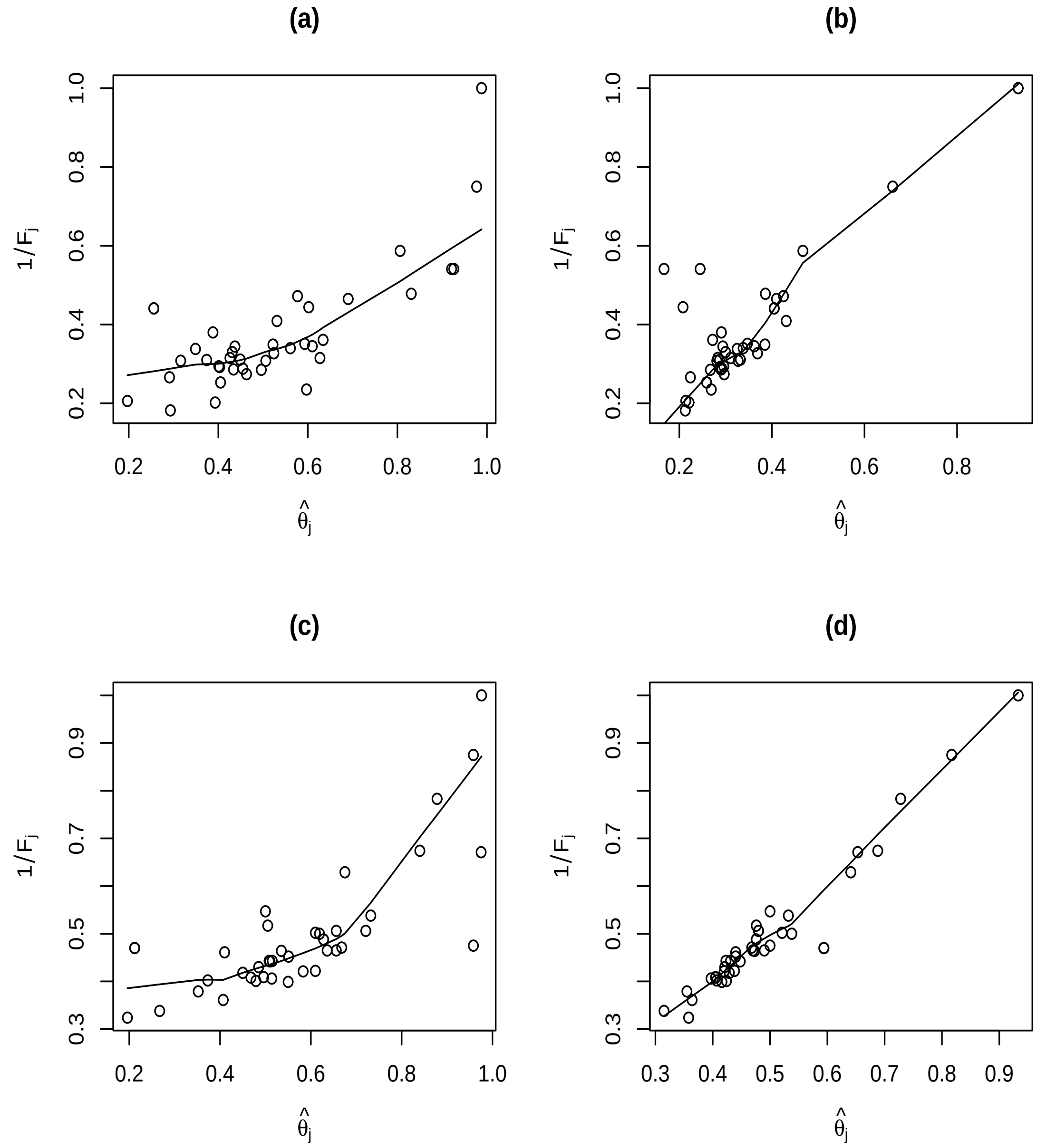

Figure 1: Scatter plot of mean of estimated measure of risk $\widehat{\theta}_j$ and mean of $1/F_j$ and loess curves with smoother span 2/3 for (a) Model 1 with $n = 2500$, (b) Model 2 with $n = 2500$, (c) Model 1 with $n = 5000$, (d) Model 2 with $n = 5000$.

summary, we suggest for use in practice the measure obtained from equation (6) 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 measure obtained from (6) ignores any error in estimating the parameters β of the log-linear model by $\widehat{\beta}$. 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 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.

References

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

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

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Clarendon: Oxford.

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

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Dale, A. and M. Elliot (2001). Proposals for 2001 samples of anonymized records: An assessment of disclosure risk. *Journal of Royal Statistical Society, Ser. A 164*, 427–447.

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

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

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

Fienberg, S. and U. Makov (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics 14*, 385–397.

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

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. and M. Elliot (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B 64*, 855–867.

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

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