Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 361–372

Estimating the Re-identification Risk Per Record in Microdata

C.J. Skinner and D.J. Holmes[1]

A measure of re-identification risk at the record level has a variety of potential uses in statistical disclosure control for microdata. The conceptual basis of such a measure is considered. The risk is conceived of broadly as the evidence in support of a link between the record and the unit in the population from which it is derived. For discrete key variables subject to no measurement error, a measure is derived which reflects the probability that the record is unique in the population. Under certain assumptions, two approaches are described for estimating this measure from the microdata. These approaches are applied to a 10% sample of microdata from the 1991 Census in Great Britain. It is found that the resulting risk measures can indeed be used successfully to establish whether sample unique records are unique in the population. The implications of these findings are discussed.

Key words: Key variable; log-linear model; lognormal distribution; population uniqueness; statistical disclosure control.

1. Introduction

Willenborg and de Waal (1996, p. 137) argue that ''in order to obtain a basis for SDC [statistical disclosure control] of microdata it is necessary to develop models for re-identification risks of individual records in a microdata file.'' This article sets out some proposals for such a development.

Per-record measures have a number of potential practical uses. Records with the highest risk might be selected for modification by SDC methods. The listing of records in order of measured risk would be analogous to similar exercises which are conducted with survey data files in which sampling weights or outlying values are ranked and listed and subsequently modified if appropriate. In addition to (or instead of) producing quantitative measures of risk per record, it may also be useful to flag records qualitatively. For example, figures in the public eye (Skinner et al. 1994) and records which are potentially spontaneously recognisable, such as someone living on the fifth floor in a rural area, might be flagged. These records might then be ''followed up'' to assess whether they really do represent an unacceptable risk. Such a practice would again be analogous to other survey data processing procedures, for example the flagging of records which fail edit checks. For any application of SDC methods it would be desirable that the number of records modified

[1] Department of Social Statistics, University of Southampton, Southampton SO17 1BJ, United Kingdom.
Acknowledgments: The Office for National Statistics supplied the anonymised 1991 Census data used in this article under contract. For the purpose of the contract, the University of Southampton is a supplier of services to the Registrar General for England and Wales and the 1991 Census Confidentiality Act applies. The authors are grateful to Angela Dale and Mark Elliot for advice about the data.

would be small enough for the effect of the modification on subsequent data analyses to be within acceptable limits.

Since the notion of re-identification refers to a microdata record, it is natural to define the risk of re-identification also at the record level. Nevertheless, this implies that the overall aim of disclosure protection can be achieved by controlling risk record by record and this does rule out some overall measures of risk. For example, suppose that the re-identification risk of a given record is defined as the probability that this record can be re-identified, and suppose that the overall risk is defined as the probability that at least one record can be re-identified. Then the overall risk is not in general a simple function of the per-record risks unless the events that different records are re-identified are independent. The latter assumption is typically implausible.

Lambert (1993) discusses some ways in which per-record risks might be combined to provide a risk measure for the whole file. One definition would be in terms of an average across records. The definition which fits in most naturally with the practical approach considered above of selecting the most extreme records is to take the overall risk as the maximum of the per-record risks. Lambert (1993) refers to this as the *pessimistic* risk.

In Section 2 we set out a broad conceptual basis for the definition of re-identification risk at the record level. In Section 3 we narrow our focus to the case where a set of discrete key variables may be available to an intruder and no measurement error may be assumed. We develop a measure of risk as the conditional probability of population uniqueness will make information available to the intruder. We base this probability on a log-linear generalisation of a Poisson lognormal model, found by Skinner and Holmes (1993) to provide an excellent fit to census microdata. In Section 4 we apply this approach to microdata on some 450,000 individuals from the 1991 Census in Great Britain and in Section 5 discuss the implications of our findings.

2. Re-identification Risk Per Record

Following Paass (1988) and Duncan and Lambert (1989), suppose that an intruder attempts to link a target person with a record on the released microdata file. The intruder may estimate the probability that each record belongs to the target person and perceive that re-identification has been achieved if this estimated probability is sufficiently large for some record. Lambert (1993) distinguishes between such ''perceived identification,'' which will depend on the intruder's prior information and beliefs, and ''true identification.'' Agencies may be expected to prefer their decisions to be as free as possible of subjective beliefs and thus to focus on the risk of true identification. Lambert (1993) and Skinner et al. (1994) define this risk in terms of the proportion of released records that can be correctly identified by a given rule. Such a frequentist definition avoids dependence on subjective factors and is the natural parameter to estimate in matching studies such as that of Blien et al. (1992). However, the definition is difficult to apply as a per-record measure because any given record would either be correctly identified by a given rule or it would not. Skinner et al. (1994) allow for some variation in the risk between individuals by defining the risk relative to some subpopulation. But this definition becomes degenerate when the subpopulation is of size one.

For this reason it seems necessary to define a per-record measure of risk relative to some

modelling assumptions, and this is what we now attempt to do. Let r denote the record for which the measure of risk is to be defined and let r^* denote the population unit (individual, household, etc.) from which the record r has been obtained. As a basic measure, we let

e_r = evidence available to an intruder in support of link between r and r^*

where the meaning of ''evidence'' is to be discussed. We assume that this evidence is conditional upon a hypothetical scenario in which an intruder either (a) takes unit r^* as a target and attempts to link this unit to its corresponding record on the file, if it exists, or (b) takes the record r and attempts to link it to a unit in the population. Note that the definition is not conditional upon either r or r^* and so applies to both cases (a) and (b). If there is more than one plausible hypothetical scenario, the risk might be defined as the maximum of the corresponding values of e_r.

One approach to measuring evidence is in terms of probability. Let $p_{rs^*} = Pr$ (record r belongs to unit s^*), where the probability is with respect to a model for the key variable values both for the released microdata file and the external data source, and with respect to the sampling scheme for the microdata sample. One measure of evidence is then to take $e_r = p_{rr^*}$, where r^* is the unit from which record r is derived. The computation of p_{rr^*} for discrete key variables will be considered in Section 3. Other approaches under different modelling assumptions are given by Paass (1988), Duncan and Lambert (1989) and Fuller (1993). An alternative measure of evidence would be to take e_r as a function of p_{rr^*}. For instance, it may be argued that there is no risk if $p_{rr^*} \leq 0.1$, there is some risk if $0.1 < p_{rr^*} \leq 0.5$ and there is high risk if $0.5 < p_{rr^*}$. In this case, e_r might be scored as follows:

$$\begin{aligned} e_r &= 2 \quad 0.5 < p_{rr^*} \\ &= 1 \quad 0.1 < p_{rr^*} \leq 0.5 \\ &= 0 \quad p_{rr^*} \leq 0.1 \end{aligned}$$

The definition of e_r as an indicator variable might also be appropriate if it flags a figure in the public eye, who might be spontaneously recognised, as noted in Section 1.

3. Measures of Risk Related to Population Uniqueness

Suppose that the intruder may use k discrete key variables $X_1, \ldots X_k$ to match microdata records with external information. Let the combinations of values of these variables in the population be denoted $x = 1, \ldots, K$ and termed *key values.* The variable taking values x is denoted X. Suppose the intruder finds that record r matches a target unit s^* with respect to X. Let F_x be the number of units in the population with $X = x$ and let $x(r)$ denote the value of X for record r. Then, assuming that $F_{x(r)}$ is known, that there is no measurement error in X (which could lead to false matches) and that microdata units are selected from the population with equal probability, the intruder may infer that record r is as likely to belong to the target unit s^* as to any of the other $F_{x(r)} - 1$ population units with this value $x(r)$ and so

Pr(record r belongs to target $s^* | F_{x(r)}) = 1/F_{x(r)}$

where the probability is evaluated with respect to the sampling scheme. In practice, the

intruder will generally be uncertain about the value $F_{x(r)}$. If the intruder may attach a probability distribution $Pr(F_x)$ to F_x then the unconditional probability is

$$e_r = Pr(\text{record } r \text{ belongs to target } s^*)$$
$$= Pr(F_{x(r)} = 1) + Pr(F_{x(r)} = 2)/2 + Pr(F_{x(r)} = 3)/3 + \ldots \quad (1)$$

This provides one definition of the re-identification risk. A simpler measure, equating population uniqueness with re-identification, would be to take

$$e_r = Pr(F_{x(r)} = 1) = Pr(\text{record } r \text{ is unique in population}) \quad (2)$$

In either case the probabilities should be evaluated conditional upon the information available to the intruder and specifically on $f_{x(r)}$, the number of sample records with key value $x(r)$. For all records which are not unique with respect to X even in the sample, that is $f_{x(r)} \geq 2$, we shall suppose the re-identification risk is acceptably small. We shall therefore focus on the conditional probability $Pr(F_{x(r)} = 1 | f_{x(r)} = 1)$ for the measure in (2). Our approach extends naturally if, instead, we wish to consider probabilities $Pr(F_{x(r)} = j | f_{x(r)} = 1)$ in (1) or the related overall measure of risk $\Sigma Pr(F_{x(r)} = 1 | f_{x(r)} = 1)$ considered by Fienberg and Makov (1998).

Sometimes it may be possible for the intruder to use external sources, such as population registers, to make an inference about $F_{x(r)}$ (Skinner et al. 1994). Here we suppose that the agency may reasonably assume that no such external evidence is available. Instead we suppose that the intruder is only able to make an inference about $F_{x(r)}$ using the released sample microdata file. We now proceed to consider how this may be done on the basis of modelling assumptions.

Following Skinner and Holmes (1993), suppose the F_x are generated independently from Poisson distributions with rates λ_x:

$$F_x | \lambda_x \sim Po(\lambda_x), \quad x = 1, \ldots, K$$

Unconditional on the key variables defining X, it seems reasonable to suppose that the λ_x are generated from a common distribution $g(\lambda_x)$. Bethlehem et al. (1990) take g to be the gamma distribution. Skinner and Holmes (1993) note that the gamma provides a poor fit and argue instead for the use of a mixture of a point mass at zero (to allow for impossible key values) and the lognormal distribution

$$\log \lambda_x \sim N(\mu, \sigma^2) \quad (3)$$

Chen and Keller-McNulty (1998) consider another distribution for λ_x. Whatever choice of g is taken, the marginal probability of population uniqueness for a unit r randomly selected from the population is

$$Pr(F_{x(r)} = 1) = P_1 / \sum_j j P_j \text{ where } P_j = \int \frac{e^{-\lambda} \lambda^j}{j!} g(\lambda) d\lambda$$

This probability is the same for all records in the file, however, and thus does not serve as a useful per-record measure. An alternative measure is proposed by Skinner et al. (1994). They consider the probability of population uniqueness amongst records which are unique in the sample microdata. To obtain an expression for this conditional probability within our framework, recall that f_x is the number of sample records with key

value x, corresponding to the population number F_x. Treating the sample as obtained via Bernoulli sampling with sampling fraction $p = n/N$, where n and N are the sample and population sizes, respectively, we may write

$$f_x|\lambda_x \sim Po(p\lambda_x) \text{ and } F_x - f_x|\lambda_x \sim Po[(1-p)\lambda_x] \quad x = 1, \ldots, K$$

where f_x and $F_x - f_x$ are independent given λ_x. It follows that the probability of population uniqueness amongst records which are sample unique is

$$Pr(F_{x(r)} = 1|f_{x(r)} = 1) = \int \exp\left[-(1-p)\lambda\right] g(\lambda|f = 1)d\lambda \tag{4}$$

where

$$g(\lambda_x|f_x = 1) = \lambda_x e^{-p\lambda_x} g(\lambda_x) / \int \lambda e^{-p\lambda} g(\lambda)d\lambda \tag{5}$$

is the conditional probability density function of λ_x given $f_x = 1$.

The measure in (4) is still constant, however, across records which are sample unique, and thus still of little use as a per-record measure. Instead, it seems desirable to condition on the values of the key variables defining X. Recall from (3) that we have assumed that λ_x is either zero or else is generated by

$$\log\lambda_x = \mu + \epsilon_x \quad \epsilon_x \sim N(0, \sigma^2)$$

This defines a log-linear model with a single intercept term and a random effect ϵ_x. We propose to generalise this model by including main effects and interactions between the key variables $X_1, \ldots, X_k$. Letting x correspond to the values $x_1, \ldots, x_k$ of $X_1, \ldots, X_k$, respectively, we may write the model including just main effects as:

$$\log\ \lambda_x = \eta_x + \epsilon_x \quad \epsilon_x \sim N(0, \sigma^2) \tag{6}$$

$$\text{where } \eta_x = \mu + u_{x_1}^{X_1} + \ldots + u_{x_k}^{X_k} \tag{7}$$

and where the $u_{x_i}^{X_i}$ represent the usual main effects of the categories x_i of X_i in a log-linear model, summing to zero for each X_i (e.g., Agresti 1990, p. 151). Some empirical evidence in support of this model is provided by Marsh et al. (1994), who demonstrated a linear relationship between the logarithm of the proportion of population uniques in a cell (corresponding to $\log\lambda_x$) and the logarithms of the univariate marginal proportions (corresponding to the $u_{x_i}^{X_i}$). Note also that if the central limit theorem applies to the sum of terms in (7) then the terms $\eta_x + \epsilon_x$ in (6) may be expected to remain approximately normally distributed, which accords with the empirical evidence of Skinner and Holmes (1993).

The model in (6) and (7) differs from a standard log-linear model only because of the term ϵ_x. Without this term the model corresponds to that considered by Fienberg and Makov (1998). The model may be extended by including terms reflecting interactions between the key variables. If enough interaction terms are included we may expect that the ϵ_x term will be unnecessary. However, the inclusion of large numbers of interactions may lead to instability in the estimation of $Pr(F_{x(r)} = 1|f_{x(r)} = 1)$. It may also be computationally more complicated and raises issues of model selection. The alternative is to include only simple terms such as main effects and to capture lack of fit by the ϵ_x terms.

Under the model defined by (6) and (7), the probability that $F_{x(r)} = 1$ given that $f_{x(r)} = 1$ retains the form in (4) but now $g(\lambda_x)$ in (5) takes the form

$$g(\lambda_x) = (2\pi\sigma^2)^{-\frac{1}{2}}\lambda_x^{-1}\exp[-(\log\lambda_x - \eta_x)^2/2\sigma^2] \tag{8}$$

This conditional probability still depends on the unknown parameters determining η_x and σ^2. A fully Bayesian approach would integrate out these parameters with respect to their posterior distribution. For simplicity, we here consider replacing the unknown parameters by point estimates. If H_i is the number of categories of X_i then the number of ''independent'' parameters μ and $u_{x_i}^{X_i}$ determining η_x in the main effects model in (7) is $J = 1 + \Sigma(H_i - 1)$. To estimate these parameters we treat (6) as an overdispersed log-linear model and estimate the mean $\mu_x = pE(\lambda_x) = p\ \exp(\eta_x + \sigma^2/2)$ of f_x under (6) in the usual way for this main effects only log-linear model (e.g., Agresti 1990, p. 170) by n times the product of the marginal proportions for which $X_i = x_i$ for $i = 1, \ldots, k$. For a more general log-linear model iterative proportional fitting (IPF) may be used. The resulting estimate is denoted $\hat{\mu}_x$. A problem is that we do not know the proportion of cells for which $\lambda_x = 0$. To allow for this in the estimation of σ^2 we consider only using data from cells for which $f_x \geq 1$. The first two moments of f_x conditional on this event are

$$E(f_x|f_x \geq 1) = \mu_x/(1 - P_{ox}),\ E(f_x^2|f_x \geq 1) = [1 + \mu_x\exp(\sigma^2)]\mu_x/(1 - P_{ox})$$

where $P_{ox} = Pr(f_x = 0)$.

Noting that

$$\frac{E(f_x^2 - f_x|f_x \geq 1)/\mu_x^2}{E(f_x|f_x \geq 1)/\mu_x} = \exp(\sigma^2)$$

we set

$$\hat{\sigma}^2 = \log\left\{\left[\sum_x (f_x^2 - f_x)/\hat{\mu}_x^2\right] \Big/ \left[\sum_x f_x/\hat{\mu}_x\right]\right\} \tag{9}$$

Note that the sums may be over all cells and not just those for which $f_x \geq 1$, since the values summed are both zero when $f_x = 0$. The per-record measure of risk is obtained from (4) and (5), where $g(\lambda_x)$ is defined in (8), η_x is replaced by $\hat{\eta}_x = \log[\hat{\mu}_x/\{\, p\exp\,(\hat{\sigma}^2/2)\}]$ and σ^2 is replaced by $\hat{\sigma}^2$ to give

$$\hat{P}(F_x = 1|f_x = 1) = \frac{\int \exp[-\lambda - (\log\lambda - \hat{\eta}_x)^2/2\hat{\sigma}^2]d\lambda}{\int \exp[-p\lambda - (\log\lambda - \hat{\eta}_x)^2/2\hat{\sigma}^2]d\lambda} \tag{10}$$

Note that this expression approaches one as $p \rightarrow 1$. The numerator and denominator of (10) may be evaluated using numerical integration as described in Skinner and Holmes (1993). In a fully Bayesian approach the support of the posterior distribution would be $[0, \infty]$. In our simplified approach, however, it is possible for $\hat{\sigma}^2$ to be negative. In this case σ^2 may be taken to be zero, so that the term ϵ_x in (6) disappears and the conditional probability in (4) reduces to

$$Pr(F_x = 1|f_x = 1) = Pr(F_x - f_x = 0) = \exp[-(1 - p)\lambda_x]$$

Estimating $\lambda_x = \exp(\eta_x)$ again by $\hat{\mu}_x/p$, the simple risk measure is given by

$$\hat{P}(F_x = 1|f_x = 1) = \exp[-(1 - p)\hat{\mu}_x/p] \tag{11}$$

Table 1. Distribution of population counts

F_x	Frequency
0	74,067
1	3,232
2	1,263
3	618
4	433
5	336
:	:

This measure might also be considered even if $\hat{\sigma}^2 > 0$ since it is much easier to compute than the measure in (10).

4. An Example: Census Microdata

We use data from the 1991 Census of Population in Great Britain on about 450,000 individuals from one local authority (see Elliot et al. 1998, and Acknowledgments). Following consideration of possible intruder scenarios by Elliot et al. (1998), we use the following $k = 6$ key variables

X_1 = age in five-year bands (19 categories)
X_2 = sex (2 categories)
X_3 = ethnic group (10 categories)
X_4 = marital status (5 categories)
X_5 = economic activity (11 categories)
X_6 = geography (4 categories)

By including variable X_6, we effectively assume that the sample microdata are released with these four geographical subdivisions identified. This corresponds roughly to the minimum area population threshold of 120,000 employed by the Census Offices in the release of anonymised individual sample microdata from the 1991 Census in Great Britain (Marsh 1993).

There are $K = 19 \times 2 \times 10 \times 5 \times 11 \times 4 = 83,600$ possible key values x defined by the combinations of values of these key variables. The frequency distribution of the population counts F_x for $x = 1, \ldots, K$ is given in Table 1. A large proportion of the key values never arise in this population. These key values with $F_x = 0$ are sometimes structural

Table 2. Distribution of sample counts

f_x	Frequency
0	79,603
1	1,585
2	531
3	339
4	220
5	156
:	:

Table 3. Percentage of population unique records re-identified by two re-identification risk measures for main effects model

Range of values of risk measure	Risk measure in (10)		Simplified risk measure in (11)	
	Number of sample unique records	Percentage population unique	Number of sample unique records	Percentage population unique
0–0.1	416	6.7	893	9.2
0.1–0.2	473	11.4	65	9.2
0.2–0.3	232	14.2	76	10.5
0.3–0.4	194	25.8	66	18.2
0.4–0.5	119	41.2	62	21.0
0.5–0.6	81	48.1	71	28.2
0.6–0.7	49	61.2	95	32.6
0.7–0.8	11	63.6	76	34.2
0.8–0.9	9	100.0	89	50.6
0.9–1.0	1	100.0	92	62.0
Total	1,585	18.9	1,585	18.9

zeros, for example children in the lowest age bands who fall into certain marital status or economic activity categories, and sometimes do not arise by chance. The number of individuals who are population unique is 3,232, representing 0.72% of the population.

In order to illustrate the approach discussed in Section 3 we drew a sample by Bernoulli sampling with sampling fraction $p = 0.1$. The achieved sample size was $n = 45,006$. The distribution of the resulting sample counts f_x is given in Table 2. The percentage of individuals in the sample who are sample unique is 3.52%. Amongst the 1,585 sample unique individuals there are 300 population unique individuals, that is 18.9%.

We then estimated the μ_x and σ^2 as described in Section 3. We first assumed the main effects model given by (6) and (7). The number of parameters determining η_x to be estimated from the one-way margins is $J = 1 + (19 - 1) + (2 - 1) + \ldots + (4 - 1) = 46$. The estimate of σ^2 from (9) was $\hat{\sigma}^2 = 3.49$.

Next we used these parameter estimates to calculate the risk measure in (10) for each of the 1,585 sample unique records. We divided the resulting 1,585 risk values according to the ranges defined in the first column of Table 3 and record in the second and third columns of Table 3 the number of sample unique records falling into each range and the percentage of these which are population unique.

A strong relationship is found between the risk measure and the proportion of population uniques within each range of values of the risk measure. The percentages increase monotonically and correspond well with the risk measures although the risk measure tends to overestimate the percentage a little away from the extremes. An intruder selecting any of the ten records with a risk measure over 0.8 would always be successful. Selecting the 70 records with a risk measure over 0.6 would lead to a success rate always over 60%. Fortunately, the majority (71%) of the records have risk measures below 0.3 for which the success rate is always under 15%.

We next recalculated the risk measure using the simplified formula in (11), which does not require numerical integration. The results are given in the final two columns of Table 3. There remains a strong relation between the percentages unique and the values of the risk measure. However, the agreement is much poorer, with the risk measure tending to overestimate the percentage population unique for all cases when the risk exceeds 0.1. The distribution of values of the simplified risk measure is also quite different. The use of the simplified measure for identifying high risk cases is not necessarily worse, however. For example, the 92 records that have the highest simplified risk measure have success rate over 60%, compared with the 70 records that have the highest values of the risk measure (10) with a success rate over 60%. Nevertheless, the value of the latter risk measure is much more accurate. For the former 92 records the simplified risk measure takes values over 0.9 overstating the actual success rate, whereas the latter 70 records have risk measures from 0.6 upwards which more accurately reflect the actual success rate.

Next we considered fitting a log-linear model including all two-factor interactions. This involved using IPF to fit the $\hat{\mu}_x$ to agree with the 987 two-way margins. Of these, 122 turned out to have sample counts of zero. All combinations x corresponding to these margins were taken as structural zeros with $\hat{\mu}_x = 0$, and IPF was applied to the remaining combinations x. The IPF algorithm reached approximate convergence after four cycles through all the margins. The estimate of σ^2 from (9) turned out to be negative, suggesting

Table 4. Percentage of population unique records re-identified by the simplified risk measure for all two-way interactions model

Ranges of values of risk measure	Number of sample unique records	Percentage population unique
0–0.1	945	3.2
0.1–0.2	137	15.3
0.2–0.3	86	29.1
0.3–0.4	52	19.2
0.4–0.5	61	31.1
0.5–0.6	57	43.9
0.6–0.7	52	55.8
0.7–0.8	71	60.6
0.8–0.9	46	63.0
0.9–1.0	78	88.5
Total	1,585	18.9

that the error terms ϵ_x in (6) are unnecessary. We therefore only consider the simplified risk measure in (11). Results corresponding to Table 3 are displayed in Table 4.

There is again a slight tendency for this simplified risk measure to overstate the percentage of population uniques, although this effect is less marked than in Table 3. This risk measure is, however, more successful than those in Table 3 in discriminating between records which are population unique and those which are not. Some 60% (945) of the records are identified as having very low risk (< 0.1) and of these only 3.2% are population unique. On the other hand, 78 records are identified as having high risk (> 0.9) and of these 88% are population unique. In comparison, of the 70 records identified by risk measure (10) for the main effects model as having the highest risk (> 0.6) only 67% are population unique.

5. Discussion

The distinction between sample uniqueness and population uniqueness is important in the assessment of disclosure risk for categorical microdata. If an intruder succeeds in matching a released microdata record to some known individual (or other unit) with respect to a combination of matching variables which can be inferred with high probability to be unique in the population, then that record has been re-identified (assuming also the absence of measurement error). On the other hand, if the intruder only knows that this record is unique in the released sample then it is possible that the record belongs to some other individual (or unit) in the population and re-identification is thus not established. Sampling can therefore be an effective means of reducing disclosure risk.

In this article we have considered how an intruder might be able to use the released microdata file to infer whether given sample unique records are also population unique (any records which are not sample unique cannot be population unique). We have proposed two measures which estimate the probability that a sample unique record is population unique: one in Equation10 which requires numerical integration and a simplified measure in Equation 11. Both measures depend on the specification of a log-linear model

for an assumed set of key variables. We have applied these measures to a 10% sample of microdata from a population of some 450,000 records from the 1991 Census of Population in Great Britain. We have found that these measures can indeed by useful in predicting population uniqueness. For example, the most successful measure, based on an all two-way interactions log-linear model, estimates that 78 records out of the 1,585 sample unique records in the sample of 45,006 records have a probability of population uniqueness greater than 0.9 and, indeed, 69 of these 78 records, i.e., 88%, turn out to be population unique. Thus, log-linear modelling does appear to have potential useful applications in statistical disclosure risk assessment. Somewhat different results were obtained for the two models considered and so results might also change with the use of more elaborate models.

These results depend on several further strong assumptions, including the following: (a) the rich set of key variables considered in Section 4 is available to the intruder for matching, (b) the sampling fraction is 10% and (c) there is no measurement error. Further research is needed to assess the effect of realistic departures from these assumptions.

Subject to such assumptions, the implication of these findings is that disclosure control methods should be applied if any records are found with high levels of risk. If only a few records of this kind are found, for example 78 constitutes just 0.2% of the sample of 45,006 records, then it is natural to consider modifying just these records, for example by replacing some key variable values by missing value codes. Considerable care should be taken with such an approach, however, since the high risk records will almost certainly be atypical and so such selective application of disclosure control methods may result in various biases. For this reason it may also be desirable to consider other more global methods such as recoding.

6. References

Agresti, A. (1990). 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, 55, 38–45.

Blien, U., Wirth, H., and Müller, M. (1992). Disclosure Risk for Microdata Stemming from Official Statistics. Statistica Neerlandica, 46, 69–82.

Chen, G. and Keller-McNulty, S. (1998). Estimation of Identification Disclosure Risk in Microdata. Journal of Official Statistics, 14, 79–95.

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

Elliot, M.J., Skinner, C.J., and Dale, A. (1998). Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. Paper presented at the Statistical Data Protection Conference, Lisbon, March.

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

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

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

Marsh, C. (1993). The Samples of Anonymised Records. Chapter 11 in A. Dale and C. Marsh (eds.) The 1991 Census User's Guide. London: Her Majesty's Stationary Office.

Marsh, C., Dale, A., and Skinner, C.J. (1994). Safe Data Versus Safe Settings: Access to Microdata from the British Census. International Statistical Review, 62, 35–53.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata. Journal of Business and Economic Statistics, 6, 487–500.

Skinner, C.J. and Holmes, D.J. (1993). Modelling Population Uniqueness. Proceedings of the International Seminar on Confidentiality, Dublin, 175–199.

Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure Control for Census Microdata. Journal of Official Statistics, 10, 31–51.

Willenborg, L. and de Waal, T. (1996). Statistical Disclosure Control in Practice. New York: Springer.

Received October 1997
Revised August 1998