Balancing Disclosure Risk Against the Loss of Nonpublication

Alan M. Zaslavsky[1] and Nicholas J. Horton[2]

A nondisclosure policy for tabular data on microdata restricts release of information that could be related to a specific individual. Pannekoek and de Waal (1998) describe a rule that suppresses data release when the number of people in a cell defined by a rare characteristic falls below a fixed floor, and show how empirical Bayes methods can be used to improve the estimation of that number. We argue that the nondisclosure problem can be formulated as a decision problem in which one loss is associated with the possibility of disclosure and another with nonpublication of data. This analysis supports a decision on whether to disclose information in each cell, minimizing the expected sum of the two losses. We present arguments for several loss functions, considering both tabular and microdata releases, and illustrate their application to simple simulated data.

Key words: Confidentiality; disclosure control; decision analysis; cell suppression; microdata.

1. To Publish or Not to Publish? That Is the Question

Statistical agencies are concerned about disclosure of confidential information when data are released that can be identified as referring to a small group of people. Disclosure can occur with tabular data releases if a cell corresponds to a very small group, for example, the cross-classification of geography with a distinctive characteristic with many levels such as occupation. Similarly it can occur in a microdata release if variables in the data can be combined with publicly available information to identify the person to whom an individual record corresponds. Again, this is particularly likely when detailed geography is combined with a characteristic like occupation, although combinations of apparently innocuous variables such as age, sex, and race may also lead to disclosure. In either case, information reported for the identified cell or microdata record (such as mean income for a cell or income for a record) can be associated with an individual or small group of individuals, violating the confidentiality of their data. Common strategies for preventing such disclosures aim to limit reporting to aggregates consisting of some minimum number of individuals, so that tabular summaries or microdata records cannot be attached to individuals or small groups of individuals. For example, cells in a table may be suppressed or combined until a fixed minimum number of cases is attained, or geographical detail may be limited to units exceeding a certain size.

This strategy is further complicated when the data are obtained from a sample survey. In that case, the number of population units represented by a particular cell may be unknown; only an approximate estimate from the sample is available. Pannekoek and de Waal (1998) address this problem where the disclosure rule is based on a minimum required population count in each of a set of predefined classes of individuals with some identifying set of characteristics. As they point out, the usual weighted estimate of the population count may be highly variable for small domains, while the synthetic estimate, obtained by multiplying the class proportion in a large area by the population of the small area, is unresponsive to local variations in the prevalence of the class. They propose using an empirical Bayes estimation procedure that combines small-domain sample estimates of prevalence of a class with rates estimated from a larger area, using relative weights that depend on the precision of each of the sources of information (Ghosh and Rao 1994).

Bayesian (or empirical Bayes) models for small-area estimation provide the posterior distribution of the population count in each domain, of which any point estimate (usually a posterior mean) is only a summary. Consequently, for each domain we can estimate the risk associated with disclosure, defined as the expectation of the disclosure loss, averaged with respect to our uncertainty about the actual population counts. The decision on whether to publish data or not is then determined by whether the risk of publication exceeds the loss for nonpublication.

In the remaining sections, we first state this approach formally (Section 2) and propose loss functions representing the risk of disclosure (Section 3). We next illustrate the calculations required to make nondisclosure decisions and show how to evaluate the sensitivity of the decisions to the choice of loss function (Section 4). We conclude (Section 5) by suggesting directions for future research.

2.   Nondisclosure as a Decision Problem

Suppose that a population is partitioned into a set of domains (such as small geographical areas or political units) and there is a class that cuts across the domains (such as people with a particular occupation). We refer to the intersection of the domain and the class as a cell. A sample is drawn from each domain, which for simplicity of exposition we assume to be a simple random sample. The population in domain i consists of N_i units of which Y_i, the cell population size, are in the class of interest. The corresponding sample contains n_i units of which y_i, the cell sample size, are in the target class, with $y_i \leq \min(n_i, Y_i)$. Of these quantities, N_i and n_i are fixed in the design, y_i is observed, and Y_i is unknown.

For each domain, a publication decision must be made for data from the target class. Again for simplicity of exposition, we assume that the only alternatives are to publish or suppress the data, although in practice there could be other options such as scrambling or rounding the data or merging cells. With these options we associate losses $L_d(y_i, Y_i, n_i, N_i)$, where $d = 1$ for publication, $d = 0$ for suppression. (For brevity, we omit some or all of the arguments of L_d when they are not required).

If L_0 or L_1 depends on the unobserved quantity Y_i, then the loss associated with the corresponding decision is unknown. If we regard the pairs (y_i, Y_i) as draws from a random

population with a known distribution, we can calculate the posterior distribution $P(Y_i|y_i)$ of Y_i. Note that $Y_i - y_i$ is the number of units in the target class among the $N_i - n_i$ non-sample unit. The risk (expected loss) for decision d is then

$$R_{di} = EL_d(y_i, Y_i, n_i, N_i)y_i = \sum_{Y_i=y_i}^{y_i+N_i-n_i} L_d(y_i, Y_i, n_i, N_i)P(Y_i|y_i)$$

(Of course, if $L_d(y_i, Y_i, n_i, N_i)$ does not depend on Y_i, we can obtain R_{di} without calculating probabilities.) The optimal decision rule is then to publish if $R_{0i} > R_{1i}$, and to suppress otherwise.

A convenient way to specify the prior distribution of (y_i, Y_i) is to suppose that the population of domain i is drawn from a superpopulation in which the probability that a unit is in the target class is θ_i, and that the θ_i for the various domains are drawn from some prior distribution. Suppose that $\theta_i \sim \text{Beta}(\alpha, \beta)$; for simplicity we assume that α, β do not depend on N_i or other covariates, although the modification for covariates is not difficult. The superpopulation sampling assumption implies that $y_i \sim \text{Bin}(n_i, \theta_i)$ and $Y_i - y_i \sim (N_i - n_i, \theta_i)$ are independent binomial draws corresponding to the sampled and nonsampled parts of the population, respectively. We describe the inference under the assumption tht (α, β) are known; in practice, they are estimated from the data but if the number of domains used in estimation is large, the variability due to estimation of (α, β) is of lower order than other sources of variability. The posterior parameter distribution is $\theta_i|y_i \sim \text{Beta}(\alpha + y_i, \beta + n_i - y_i)$. Integrating over this distribution, the predictive distribution of $(Y_i - y_i)|y_i$ is the beta-binomial (Polya Type I) distribution with density

$$P(t; \alpha^*, \beta^*, N^*) = \binom{N^*}{t} \frac{\Gamma(\alpha^* + \beta^*)\Gamma(t + \alpha^*)\Gamma(N^* - t + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)\Gamma(\alpha^* + \beta^* + N^*)}$$

where $\alpha^* = \alpha + y_i$ and $\beta^* = \beta + n_i - y_i$ are the parameters of the Beta prior, $N^* = N_i - n_i$ is the (nonsampled) population size, and $t = Y_i - y_i$ is the value of the random variable. The distribution can be readily enumerated to calculate R_{0i}, R_{1i}. The posterior expectations of both θ_i and $(Y_i - y_i)/(N_i - n_i)$ are $(\alpha + y_i)/(\alpha + \beta + n_i)$, so that posterior expectation of Y_i is $y_i + (N_i - n_i)(\alpha + y_i)/(\alpha + \beta + n_i)$.

In practice, there may be rough consensus on the relative losses for nonpublication of different cells, and likewise for the relative losses for potential disclosure in different cells, but not consensus on the relative importance of publication and nondisclosure. If L_0 and L_1 are known only up to a proportionality factor, we can still specify a set of decision rules that minimize the risk at various values of that factor. These rules are of the form ''publish if $R_{1i}/R_{0i} < c$.'' Equivalently, we can choose to publish a predetermined fraction of the data by ordering cells by the ratio R_{1i}/R_{0i} and publishing from the top of the list until the target is attained; this allows us to compare the cell publication decision under rules based on different loss functions that are not comparable.

3.   Loss Functions for Suppression and Disclosure

The cost L_0 of suppression of a cell is that the public is deprived of potentially useful information. We may regard loss as equal for each cell that is suppressed, or relate loss to some measure of the size of the cell. The number of microdata records that are concealed is y_i.

We may also define loss by the number of population units that they represent, Y_i, or by sample or population counts for the entire domain, n_i or N_i.

The loss L_1 represents the potential disclosure consequences of publication of a cell. In a microdata release, or in a tabular release where cells are further cross-classified with other variables, we would be concerned about the possibility of identification of an individual by matching of characteristics in a microdata record or cross-classified cell to publicly known characteristics of some person. Beyond this, we may be concerned about release of data (such as mean income) for a very small group even if individuals cannot be identified (Duncan and Lambert 1986). Even the perception that disclosure is possible may impose some costs (Lambert 1993).

In either case, typically L_1 would be decreasing in y_i and/or Y_i because reporting on a small group presents a high risk of disclosure. A reasonable requirement is that L_1 is convex, i.e., that the reduction in loss of each additional unit in the cell is decreasing, meaning that the protection against disclosure for each additional person is less when the cell is already large.

If the identities of the sampled respondents are known to a person attempting to invade privacy, then the risk of disclosure depends on y_i, the number of those respondents in the cell.

More commonly, as in Pannekoek and de Waal's approch, the identity of the sample is confidential and only the members of the population of the cell are publicly known. In this case, the loss is decreasing in Y_i, and may be reasonably assumed to be convex for the reasons given above. For a microdata release, the loss is also approximately proportional to y_i because the risk of disclosure increases with the number of records that are potentially reidentifiable.

To clarify the meaning of the loss function, consider the rule used in the Netherlands and reported by Pannekoek and de Waal, ''publish if $Y_i > Y_{min}$'' (for some constant Y_{min}). This is a fairly good decision rule, especially if sample and population sizes do not vary much across domains. On the other hand, ''$L_1 = 1$ if $Y_i < y_{min}$, $L_1 = 0$ otherwise'' is not a very sensible loss function, because it implies that any $Y_i < Y_{min}$ poses an equal danger of disclosure.

We may regard each class as divided into subclasses consisting of sets of individuals who can be distinguished using published microdata fields or crosstabulation categories; Greenberg and Zayatz (1992) refer to these as ''equivalence classes'' because individuals in them are equivalent from the point of view of identifiability. From this perspective, a person's data are disclosed if the person is in the sample and the sole member of her subclass within the domain, i.e., if a ''sample unique'' is also a ''population unique'' (Samuels 1998). We can then quantify loss as the expected number of individuals who will be disclosed in this sense, which depends on y_i, Y_i and the distribution of the class across subclasses. A substantial literature estimates this quantity through probabilistic analyses; see Chen and Keller-McNulty (1998) for a bibliography, and Samuels (1998), Fienberg and Makov (1998), and Skinner and Holmes (1998) for current theoretical and empirical investigations. We illustrate such analyses here by an argument along the lines of Bethlehem, Keller, and Pannekoek (1990) or Skinner, Marsh, Openshaw, and Wymer (1994), although this particular model has been found not to fit well in empirical analyses.

Suppose that a class consists of S subclasses, and the fraction of the population in

subclass s is π_s, $0 < \pi_s \ll 1$. If a sample person is in subclass s, the probability that there are no other people in subclass s in the cell is approximately (using a Poisson approximation to the binomial) $\exp(-Y_i \pi_s)$. Unconditionally, the probability that a particular sample case is unique is $\Sigma_s \pi_s \exp(-Y_i \pi_s)$, so the expected number of uniques in the sample is

$$L_1 \approx y_i \sum_s \pi_s \exp(-Y_i \pi_s) \tag{1}$$

If all the classes are of the same size, $\pi_s = 1/S$, then

$$L_1 \approx y_i \exp(-Y_i/S) \tag{2}$$

If the class sizes vary and the π_s have an approximate Gamma (a, b) distribution, then $E\pi_s = a/b$ so $Sa/b \approx 1$. Evaluating (1) as an integral with respect to the distribution of π_s, we obtain $L_1 \approx y_i/(1 + Y_i/aS)^{(a+1)}$. As $a \to \infty$, indicating an increasingly peaked distribution of subclass sizes, this approaches (2). These loss functions possess the properties of monotonicity and convexity suggested above.

The characteristics of loss functions described above have implications for the decision analysis. By Jensen's inequality and the convexity of the loss function, $R_i = E L_1(y_i, Y_i) > L_1(y_i, E Y_i)$, where the expectation is with respect to the posterior distribution of Y_i. In other words, plugging in the posterior expectation of Y_i underestimates loss. On the other hand, for a smooth loss function, the Taylor series approximation to the risk is $R_i \approx L_1(y_i, E Y_i) + (L_1''(y_i, E Y_i)/2)\text{Var } Y_i$ (where the derivative is with respect to Y_i). If $\text{Var } Y_i$ is approximately a function of $E Y_i$, as in a nearly balanced design (i.e., one in which domain sizes and sampling rates do not vary much), then R_i is also approximately a function of $E Y_i$. In that case any function of Y_i, or just $E Y_i$ itself, gives approximately the same ordering of loss by cell. On the other hand, we might expect to find larger differences among publication decisions under different loss functions if the loss functions have very different curvature and the design is very unbalanced so cells with similar values of $E Y_i$ have substantially different values of $\text{Var } Y_i$.

4. Illustrations

We first illustrate the calculations described in Section 2 using a small example, assuming the beta-binomial model described in that section. In our example, $\alpha = 1$ and $\beta = 10$. Domains are of two types by population and sample size. Type 1 with $n_1 = 3$, $N_1 = 8$, and Type 2 with $n_2 = 5$, $N_2 = 20$, appearing in a 1:3 proportion. Disclosure loss is $L_1(y_i, Y_i, n_i, N_i) = y_i \exp(-Y_i/10)$, corresponding to the expected number of disclosures with 10 equally prevalent subclasses, and nonpublication loss is $L_0(y_i, Y_i, n_i, N_i) = y_i$, the number of sample cases that are suppressed.

Table 1 displays the calculation of $R_1(y)$ for Size 1 domains in this example. The posterior probabilities $P(Y|y)$ (based on the beta-binomial distribution of $Y - y$) are multiplied by loss L_1 and the products are summed to calculate the risk for each y.

Table 2 displays all possible configurations of the observed quantities y_i, n_i, Y_i, N_i. In the fifth line of the table (just above the divider), for example, $P(n, N) = .75$ represents the fraction of domains with $n = 5$, $N = 20$, $P(y, N) = .018$ is the (beta-binomial) probability that each of those domains has $y = 3$, $R_1 = 1.565$ is obtained by a calculation

Table 1.   Calculations of risks $R_1(y)$ for $n = 3$, $N = 8$

y		Y									
		0	1	2	3	4	5	6	7	8	$R_1(y)$
0	$P(Y\mid y)$	0.722	0.212	0.053	0.011	0.002	0.000	0	0	0	
	$L_1(y, Y)$	0	0	0	0	0	0	0	0	0	
	$P \cdot L_1$	0	0	0	0	0	0	0	0	0	0
1	$P(Y\mid y)$	0	0.510	0.319	0.127	0.036	0.007	0.001	0	0	
	$L_1(y, Y)$	1	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.407	
	$P \cdot L_1$	0	0.461	0.261	0.094	0.024	0.004	0	0	0	0.846
2	$P(Y\mid y)$	0	0	0.350	0.350	0.200	0.077	0.019	0.002	0	
	$L_1(y, Y)$	2	1.810	1.637	1.482	1.341	1.213	1.098	0.993	0.899	
	$P \cdot L_1$	0	0	0.574	0.519	0.269	0.093	0.021	0.002	0	1.479
3	$P(Y\mid y)$	0	0	0	0.234	0.334	0.257	0.128	0.041	0.007	
	$L_1(y, Y)$	3	2.715	2.456	2.222	2.011	1.820	1.646	1.490	1.348	
	$P \cdot L_1$	0	0	0	0.519	0.671	0.467	0.211	0.061	0.009	1.939

like that in Table 1, and $L_0 = 3$ is the loss for suppressing a cell with 3 sample cases. We then calculate $P \cdot R_1 = 0.75 \times 0.018 \times 1.565$, the contribution to expected risk if we publish domains with $n = 5$, $N = 20$, $y = 3$ and similarly $P \cdot L_0 = 0.75 \times 0.018$, the contribution if we do not publish.

The lines of the table are ordered by the ratio R_1/L_0, because we first publish the cells with the least disclosure risk per unit of information. Our optimum rules correspond to publishing data for all configurations down to a certain line and suppressing data for configurations below that line. For example, the cumulative sum 0.026 is the expected disclosure risk per domain if we publish data for all domains corresponding to the first five lines and 0.357 is the expected nonpublication loss per domain for suppressing data for the remaining domains.

Note that with these loss functions, $R_1/L_0 = E \exp(-Y_i/10)\mid y_i$, which by the argument in the previous section is approximately a function of $E Y_i \mid y_i$. In fact the ordering of

Table 2.   Calculation of optimal publication ordering and of disclosure risk and nonpublication loss at each cutoff

n	N	y	$P(n, N)$	$P(y\mid n, N)$	R_1	L_0	R_1/L_0	$P \cdot R_1$	$P \cdot L_0$	$\sum_{pub} P R_1$	$\sum_{\overline{pub}} P L_0$
3	8	0	0.25	0.769	0.000	0	–	0.000	0.000	0.000	0.409
5	20	0	0.75	0.667	0.000	0	–	0.000	0.000	0.000	0.409
5	20	5	0.75	0.0003	1.783	5	0.357	0.0004	0.001	0.0004	0.408
5	20	4	0.75	0.003	1.726	4	0.432	0.004	0.010	0.005	0.398
5	20	3	0.75	0.018	1.565	3	0.522	0.022	0.041	0.026	0.357
5	20	2	0.75	0.073	1.261	2	0.630	0.069	0.110	0.096	0.247
3	8	3	0.25	0.003	1.939	3	0.646	0.002	0.003	0.097	0.244
3	8	2	0.25	0.035	1.479	2	0.739	0.013	0.017	0.110	0.227
5	20	1	0.75	0.238	0.761	1	0.761	0.136	0.179	0.246	0.048
3	8	1	0.25	0.192	0.846	1	0.846	0.041	0.048	0.287	0.000

configurations for suppression based on $\exp(-(E\,Y_i|y_i)/10)$ is identical to that obtained using R_1/L_0. Hence Pannekoek and de Waal's procedure works well in this situation, as might be expected from the argument at the end of Section 3.

Our second example is designed to illustrate sensitivity of publication decisions to the choice of loss function. We assume that domains fall into one of three equally-common size categories, with sample sizes $n = 10, 10, 20$ and populations $N = 50, 200, 30$, and that $\alpha = 1/2$ and $\beta = 3/2$. As before, we defined $L_1 = y_i \exp(-Y_i/10)$ and $L_0 = y_i$. The tradeoff of disclosure risk and nonpublication bias under the optimum disclosure rules is represented by the solid curve in Figure 1, the lowest that can be attained with any non-disclosure rule based on y, n, and N. This line corresponds to the plot of the next to last column of Table 2 (recalculated for the new parameter values) on the vertical axis, against the last column on the horizontal axis (both normalized to run from 0 to 1). The optimal rule substantially improves on random suppression of domains, represented by the straight dashed line. For example, with the optimal rule, we can cut disclosure risk to 20% of its value under full publication, while suppressing publication of only 38% of the data.

To check the sensitivity of the decision rule to the choice of loss function, we considered several alternative pairs of loss functions, each of which generates a corresponding decision rule. We evaluate each of the alternative decision rules in terms of the original loss functions. The alternatives are represented by the various dashed curves in Figure 1. In

Legend:
- $L_0 = y$, $L_1 = y \exp(-Y/10)$
- $L_0 = $ equal, $L_1 = y \exp(-Y/10)$
- $L_0 = y$, $L_1 = \exp(-y/10)$
- $L_0 = y$, $L_1 = y \exp(-Y)$
- Random suppression

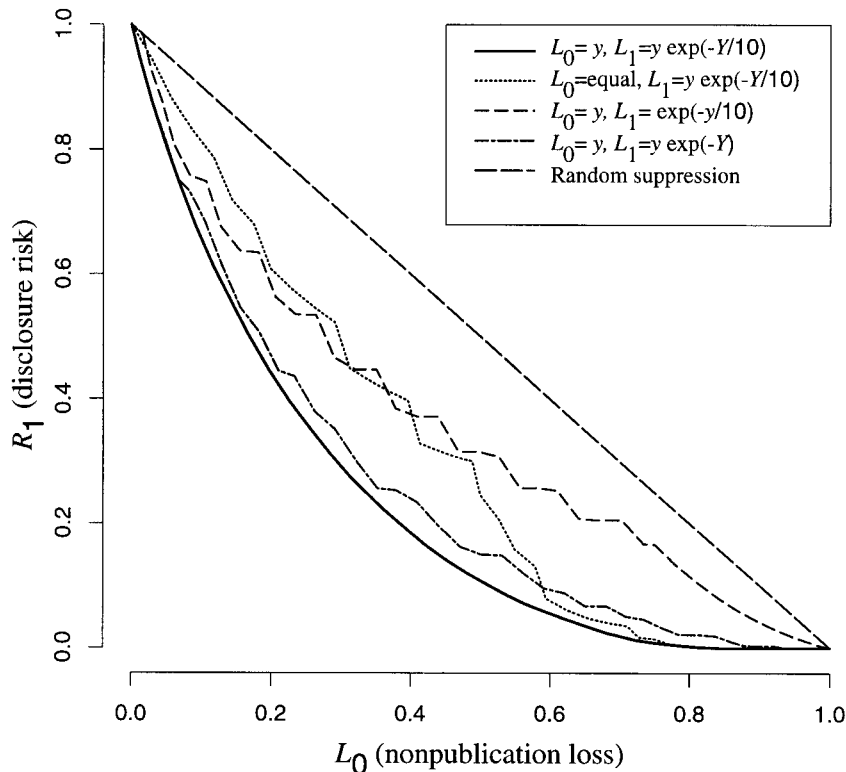Axes: R_1 (disclosure risk) versus L_0 (nonpublication loss)

Fig. 1. Disclosure risk plotted against nonpublication loss (both relative to largest possible values) for scenario described in text. The loss functions defining the axes are $L_0(y_i, Y_i) = y_i$ and $L_1(y_i, Y_i) = y_i \exp(-Y_i/10)$.

the first alternative, L_1 remains the same but $L_0 = 1$, i.e., suppression of any cell engenders the same nonpublication loss. In the second, $L_o = y$ but $L_1 = -\exp(y/10)$, i.e., decisions are taken using a rule appropriate to disclosed respondents when actually the respondents' identities are concealed. In the third, $L_1 = y \exp(-Y)$, which is similar in form to the ''correct'' L_1 but assumes the wrong number of subclasses.

In this (admittedly artificial) example, use of the ''wrong'' L_0 *or* L_1 can lead to substantially suboptimal nonpublication decisions. For example, to reduce disclosure risk to 20% of its maximum value, the optimum rule requires suppression of 38% of the data, but the corresponding percentages for the three alternative scenarios are 53%, 71%, and 44%. This illustrates that in some scenarios, serious analysis of potential disclosure (along the lines of the research referenced above) and elicitation of preferences is necessary to gain the full benefits of the decision framework.

5.   Conclusion

Like any broad conceptual framework, the application of decision analysis to the non-disclosure problem is the beginning rather than the end of a research program. Among the research tasks involved in making concrete our general proposals are: (1) defining the classes, domains, and identifiable subclasses for data sets of interest; (2) specifying numerical values L_1; (3) eliciting consensus values for data availability and nondisclosure, reflected in relative magnitudes assigned to L_0 and L_1; (4) more realistically describing decision alternatives for nondisclosure, such as combining cells or suppressing some data fields; and (5) fitting the required small-area models, as in Pannekoek and de Waal (1998). Nonetheless, a unified conceptual framework is the first step toward developing the tools required to design sound policies.

6.   References

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

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

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited Data Dissemination. Journal of the American Statistical Association, 81, 10–18.

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

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55–76.

Greenberg, B.V. and Zayatz, L.V. (1992). Strategies for Measuring Risk in Public Use Microdata Files. Statistica Neerlandica, 46, 33–48.

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

Pannekoek, J. and de Waal, A.G. (1998). Synthetic and Combined Estimators in Statistical Disclosure Control. Journal of Official Statistics, 14, 399–410.

Samuels, S.M. (1998). A Bayesian, Species-Sampling-Inspired Approach to the Uniques

Problem in Microdata Disclosure Risk Assessment. Journal of Official Statistics, 14, 373–383.

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

Skinner, C.J. and Holmes, D.J. (1998). Estimating the Re-Identification Risk Per Record in Microdata. Journal of Official Statistics, 14, 361–372.