Journal of Official Statistics
Vol. 10. No. 1, 1994, pp. 31–51

Disclosure Control for Census Microdata

Chris Skinner[1], *Catherine Marsh*[2], *Stan Openshaw*[3], *and Colin Wymer*[4]

Abstract: Approaches to disclosure control for microdata from population censuses in different countries are described. A framework for assessing disclosure risk is presented. The estimation of this risk is considered in the context of census microdata. Implications for disclosure control are discussed.

Key words: Disclosure control; disclosure risk; identification; population uniqueness.

1. Introduction

The release of population census microdata for secondary analysis can become an area of conflict. Census authorities are increasingly squeezed between users on the one hand who want more and more detailed microdata to be made available, and "public concern" about the risk of disclosure on the other hand.

Countries have responded to this dilemma in different ways. Some, such as the United States, Canada and Australia have come down in favour of releasing files of suitably anonymised census records in a formal and public manner. Others, such as Sweden and Italy have struck a balance between concern with disclosure and demand for the data by providing limited access to census microdata to selected academics and others outside the census office. These users are required to take special measures to preserve confidentiality, and may be required to do their work in the census office.

Elsewhere, however, census microdata are not available. While in most countries this may be because of lack of user demand, in some cases (e.g., in the Federal Republic of Germany), the census authorities have apparently taken the view that the risks of breach of confidentiality outweigh the benefits that would arise from making samples of census records, however anonymised, publicly available. Although no breach of confidentiality seems ever to have occurred in those countries where census microdata have been made available, some census authorities presumably fear

[1] Department of Social Statistics, University of Southampton, Southampton SO9 5NH, U.K.
[2] Census Microdata Unit, University of Manchester, Manchester, M13 9PL, U.K. Catherine Marsh died from cancer on January 1, 1993. She played a vital role in bringing about the first release of microdata from a British Census.
[3] Department of Geography, University of Leeds, Leeds LS2 9JT, U.K.
[4] Centre for Urban and Regional Development Studies, The University, Newcastle upon Tyne NE1 7RU, U.K.
Acknowledgements: The authors gratefully acknowledge the assistance provided by Fabio Sforzi of IRPET and also of ISTAT. The data used in Table 1 were supplied to Newcastle University under a joint programme of research with IRPET, Firenze. Research by C.J. Skinner was partly conducted at Iowa State University, supported by JSA 90-7, U.S. Bureau of the Census and Co-operative Agreement 68-3A75-0-9, Soil Conservation Service, U.S.D.A. The authors are grateful to the referees and the Associate Editor for their helpful comments.

that it could still happen. Official exercises like the census can easily become the symbolic focus for protests of a more general kind. As experience in Germany has shown, serious public misgivings about the use to which census data will be put can actually jeopardise the census enterprise itself (Butz 1985).

In Great Britain the Census Offices will in 1993 release two samples of anonymised records (a 2% sample of individuals and a 1% sample of households) from the 1991 Census of Population, in response to a request from the Economic and Social Research Council, on behalf of academic social scientists. The request, which is set out in a report by the present authors and others (Marsh et al. 1991), assesses the potential benefits of census microdata and sets them against the possible costs in terms of the risks of disclosure. No such microdata samples have been released before from a census in Great Britain.

The aim of this paper is to develop a general framework, within which disclosure risk can be assessed and controlled for census microdata. Some approaches to estimating the different components of disclosure risk will be surveyed. Particular reference will be made to the work of Bethlehem, Keller, and Pannekoek (1990), who showed how "population uniqueness" can be estimated from sample data if a Poisson-gamma model is assumed. Some new empirical evidence on the goodness of fit of this model will be presented.

We concentrate on the real risks of disclosure, rather than on other risks which might be perceived to exist by the public, but which are not in fact real (Courtland 1985; Cox, McDonald, and Nelson 1986). We are concerned only with the possibility of statistical disclosure, whereby a user links a microdata record to an identifiable individual via their profile of census characteristics, and not with other forms of disclosure, such as might arise from breaches of physical security (Courtland 1985; Cox et al. 1986). Moreover, we consider only methods of disclosure control which involve various kinds of statistical transformation of the raw census data to produce the microdata for public release (Dalenius 1977a) and not other contractual or administrative methods (Gates 1988).

Two broad types of disclosure control methods may be distinguished:

1. *Methods which preserve the integrity of the data*
   Examples include sampling individuals by some representative scheme, suppressing variables or combining categories of variables.
2. *Contamination methods*
   Examples include the creation of synthetic microdata records from census records for more than one individual (Paass 1988), the addition of noise (Spruill 1983; Kim 1986; Sullivan and Fuller 1989), data swapping (Spruill 1983) or nonrepresentative sampling, for example, of those individuals who do not possess unique census records (U.S. Department of Commerce 1978, p. 29; Dalenius 1986).

For multipurpose datasets such as census microdata, methods of the first type are generally preferable to users. For this reason, we devote the majority of this paper to assessing disclosure risk and its dependence on factors which can be controlled by methods of type (1). We return briefly to methods of type (2) in Section 7. In Section 2 we shall introduce our definition of disclosure risk in terms of an identification rule, which links records and verifies the link is correct. Aspects of such rules are considered in greater detail in Sections 3 and 4. The estimation of

disclosure risk is considered in Section 5 and further issues are discussed in Sections 6 and 7.

2. Assumptions, Definitions and Basic Framework

Our basic framework follows that of many authors, for example, Paass (1988), Duncan and Lambert (1989) and Bethlehem et al. (1990). We consider an **investigator** who attempts to disclose information about identifiable individuals. We assume the investigator has access to **prior information** about a set of **target individuals**, whose **identities** (names and addresses, say) are known to the investigator. This set of individuals is termed the investigator's **circle of acquaintances** by Willenborg, Mokken, and Pannekoek (1990). In order to achieve disclosure, the investigator is assumed to attempt to link the prior information for the target individuals to the microdata records, using the values of a set of **key variables** (Bethlehem et al. 1990), which are available in both the prior information and the microdata. We write $\mathbf{x}_i$ as the vector of values of the key variables in the prior information and $\tilde{\mathbf{x}}_i$ as the vector of values in the microdata for individual i. The distinction between $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ allows for the possibility of measurement error. Thus for an individual who is in both the circle of acquaintances and the microdata sample it will only be the case that $\mathbf{x}_i = \tilde{\mathbf{x}}_i$ if all the key variables are recorded in an identical way in both datasets.

While there seems to be broad agreement in the literature on the basic framework outlined above, differences do emerge in the definition of statistical disclosure. Fellegi (1972, p. 8) takes disclosure to arise if "information can be deduced from the published estimates that can be related to a particular identifiable respondent." U.S. Department of Commerce (1978) and Duncan and Lambert (1986) relax this definition by replacing the exact deterministic idea of deduction by ideas of approximate determination or of probabilistic inference. Such definitions have natural applications to disclosure control for tabular census data. For microdata, however, a common overriding concern in legislation governing the release of statistical information is to ensure that no microdata record can reliably be associated with an identifiable individual (Paass 1988). Our definition, which follows, embodies this idea. The definition is based on those of Cox and Sande (1979), and Paass (1988). A linkage is said to result in **disclosure** if both of two steps occur:

i. **identification**, i.e., the investigator succeeds in linking an individual to a microdata record and is able to verify with high probability that this link is correct;
ii. the investigator consequently obtains **new information** about this individual, which was not available in the prior information.

The requirement in (i) that the link be verified is essential since, otherwise, any guessed link could qualify as disclosure and disclosure control would be impossible. For further discussion of alternative definitions of disclosure see Dalenius (1977b), Duncan and Lambert (1989), and Skinner (1992).

The crucial part of our definition is step (i). If the prior information on a linked individual includes all the information on the microdata record then step (ii) and hence disclosure is impossible. Thus it is sufficient to consider the case where the prior information does not include all the information on the microdata record. But in this situation if step (i) can be established

then step (ii) follows automatically, provided the microdata record has not been distorted by error. In what follows we therefore focus on assessing the risk of identification.

In order to arrive at a definition of identification risk, it is first necessary to consider the reasons which might lead to an attempt at disclosure. There appear to be two possibilities:

i. There might be genuine advantages to be obtained from systematic record linkage. For example, it is at least conceivable that a commercial company which owned a large database might seek to enhance the information in it by attaching census information to any individuals it could identify.
ii. Since the census is an important symbol of government, an investigative journalist or politically motivated individual might try to disclose census information on a one-off basis purely to demonstrate that it was possible.

Given these reasons for attempting disclosure, there seem to be two possible scenarios, according to which the attempt might take place:

a. the investigator first selects some target individuals and then attempts to link the record of prior information on each of these individuals to one of the microdata records;
b. the investigator first selects one or more records from the microdata and then attempts to link each of these to individuals in the general population.

As discussed by U.S. Department of Commerce (1978, p. 27) and Paass (1988), investigators whose reason for attempting disclosure is (i), would be expected to employ scenario (a), whereas those with reason (ii) might employ either (a) or (b), in either case possibly carrying out a "fishing expedition" for records with "recognizable" combinations of characteristics. We shall first examine the risk of identification under scenario (a) and then in Section 6 assess how this risk is modified for scenario (b).

A procedure by which an investigator attempts to link a microdata record with a target individual and to verify the correctness of the link will be called an identification rule. The possible form of such rules will be considered further in Section 3. Like Duncan and Lambert (1989), we suppose that the outcome of any identification rule is that the investigator either decides to link an individual to one microdata record or else decides that there is not enough information to link the target individual to any record. The target individual will be said to be identified by an identification rule if the rule leads the investigator to link the individual to a microdata record. The identification risk for individual i and a given identification rule r is then the probability that the individual is identified:

$$\begin{aligned}&\text{Identification risk for individual } i \\ &\quad \text{and rule } r \\ &= \Pr(i \text{ identified by rule } r).\end{aligned} \tag{1}$$

The interpretation of this probability statement requires some clarification. Duncan and Lambert (1986; 1989) adopt a Bayesian approach in which probabilities correspond to the investigator's subjective beliefs. Here we adopt a frequentist approach. Under scenario (a) above, we suppose that individual i has been selected randomly from a subpopulation S and is equally likely to be any member of S. We may then equate the probability of identification to the proportion of members of S who would be identified by the rule. The nature of S may

reflect the degree of "selectivity" in the attempt at disclosure. Thus, for general attempts of type (i) above, S might consist, for example, of all adults with a given credit card, whereas for more selective attempts of type (ii), S might consist of all individuals falling into certain rare categories on one or more of the key variables.

An advantage of our definition of probability is that it may be directly estimated from empirical experiments, such as those of Blien, Wirth, and Müller (1992). In principle, the definition could be extended to a model-based approach in which the values of the key variables were generated from some model. Both in such a superpopulation framework or in our original population framework, the properties of an identification rule can be calibrated by frequentist criteria, even though the rule may itself employ Bayesian methods for verification.

The outcome of any identification rule can be either correct or incorrect and we define the rate of false positives as

$$\begin{aligned} R_{ir} &= \text{Rate of false positives of rule } r \\ &\quad \text{for individual } i \\ &= \Pr(\text{incorrect link} | i \text{ identified by} \\ &\quad \text{rule } r), \end{aligned} \tag{2}$$

where the probability statement is interpreted as before. Given the verification requirement of the definition of identification, we may restrict attention to the class $C_{i\alpha}$ of identification rules for which the rate of false positives is below some threshold α, say 10%. An identification rule might be viewed as optimal for a given individual if it maximises Pr (i identified by rule r) within this class. The overall identification risk for individual i may then be taken as

$$\begin{aligned} &\Pr(i \text{ identified}) \\ &= \max_{r \epsilon C_{i\alpha}} \Pr(i \text{ identified by rule } r). \end{aligned} \tag{3}$$

In many circumstances it will be impossible to establish a reliable link and the set $C_{i\alpha}$ will be empty, in which case Pr (i identified) may be taken to equal zero.

For simplicity, the probability of identification will be evaluated conditional on identification being attempted for just one specific target individual and conditional on a given scenario of an attempt at disclosure. The way in which such probabilities for different individuals in the investigator's circle of acquaintances might be combined into the overall probability that at least one individual is identified is discussed by Willenborg et al. (1990).

3. Identification Rules

In order to assess the risk of identification it is necessary to consider the possible form of identification rules. The simplest approach is to separate the functions of linkage and verification. A simple linkage rule involves claiming a link if the key variable values for the individual match those on a microdata record exactly. A naive verification rule associated with this linkage rule would be to claim that a link is correct when there is only one record in the microdata sample which matches the individual exactly, that is, the record is sample unique. This rule is inadequate, however, since if there is another individual in the population with the same key variable values then there is a probability of at least 0.5 that the match is incorrect (Duncan and Lambert 1989). Thus, in order to obtain an identification rule with an acceptable level of false positives, it is necessary to verify that the individual is population unique. Possible ways in which this might be achieved will be discussed in Section 4.

Provided there is no measurement error, population uniqueness will be a sufficient

condition for an exact match to be verified as correct. However, if there is measurement error then different individuals might be erroneously matched. Hence some means of verifying that this is not the case is needed. Some approaches to this problem have been suggested (Paass 1988; Duncan and Lambert 1989; Sullivan and Fuller 1989), although these have mainly dealt with continuous variables and there is not the space to pursue these ideas further here. More fundamentally, the possibility of measurement error suggests that the restriction of claimed links only to the subset of cases where there are exact matches might lead to too many false negatives and hence to a non-optimal identification rule in the sense of Section 2. A variety of linkage rules which permit discrepancies in some key variable values have been suggested. Blien et al. (1992) suggest defining alternative sets of key variables and making a link if there is an exact match with respect to any of these alternative sets. Spruill (1983) and Strudler, Oh, and Scheuren (1986) suggest linking the individual to that record which minimises the sum of absolute deviations or squared deviations over all key variables. Verification procedures for both these approaches remain unclear, however. The more sophisticated approaches of Paass (1988), Duncan and Lambert (1989) and Sullivan and Fuller (1989) incorporate the linkage rule and verification rule in one procedure by specifying a measurement error model and linking only records for which the estimated probability of a correct link is above some threshold.

The question as to whether such rules would ever lead to identification in the practical context of census microdata and what the actual rates of false positives would be is an empirical one for which at present there is only very limited evidence. In an interesting study Blien et al. (1992, p. 71) attempted to link 169,368 records from the 1987 "microcensus" in one German state to information on 10 key variables for 7,983 individuals in a handbook of German scientists and scholars. This number of key variables was "the highest number of key variables of the relevant handbooks." Using exact matching, subject to 16 alternative definitions of the key variables, and rejecting any match which was not unique in both the microcensus and the handbook, they were able to link 14 records of which only 4 turned out to be correct. The rate of false positives was therefore 10/14. They also applied Paass's (1988) method which allows for measurement error to a set of key variables for which exact matching had produced 7 links, of which 3 were correct. Under two alternative assumptions about the sizes of the measurement errors, Paass's method produced 9 or 11 links, of which only 2 were correct in each case. The rate of both false positives and false negatives was therefore worse than for the exact matching procedure. The practical failure of Paass's procedure here may be due to its lack of robustness to assumptions about the nature and magnitude of the measurement error. It might be possible to reduce the rate of false positives by reducing the threshold specified in the method but this might tend also to reduce the number of correct positives towards zero. Indeed Paass (1988) found in his study that for the only two scenarios which involved solely categorical key variables, the "address-list broker" scenario and the "industrial-enterprise" scenario, his method gave no links at all. The method only led to successful links for three scenarios all involving some continuous financial key variables which are not characteristic of census microdata.

For the remainder of this paper we

propose for simplicity to restrict attention to identification rules based on linkage rules employing exact matching. While it may be feared that such rules are non-optimal, as mentioned earlier, evidence from Blien et al. (1992) suggests that methods which allow for measurement error do *not* in fact appear to lead to more correct links for the kinds of categorical key variables and measurement error typical of census microdata. Hence, on this evidence, it appears our approach remains conservative.

4. Verification of Population Uniqueness

In Section 3 we noted that an identification rule should verify that an individual is population unique, that is, that the individual's combination of key variable values is unique in the population. In this section we discuss ways in which this verification might take place. By attempting to make each of these ways difficult for the investigator, the census authority can help to control disclosure.

One can imagine three ways in which verification might be attempted.

4.1. Population lists

If the investigator had access to some comprehensive list of the population or some specific subgroup of the population defined by a census variable, it would be possible to check whether a unique case in the sample was unique in the population (U.S. Department of Commerce 1978, p. 26). For example, if the investigator's circle of acquaintances includes a black female judge, aged 45, living in a particular area, and if the census microdata sample also yielded such a case, a list of all judges which gave age, sex, ethnicity, and area of residence could establish that such a person was unique, and thereby reveal that person's identity.

The existence or otherwise of comprehensive lists containing sufficient census variables to be usable for matching purposes is a matter for empirical enquiry in any country that is considering releasing census microdata. In Britain, it is doubtful whether any such list exists which could be used in this way. Lists of professionals holding particular qualifications usually only contain name and sex, not age, and often not even area of residence. "Yellow page" telephone directories suffer similar problems. Although several commercial companies have the whole of the electoral roll on file, the only variables they can extract from that are area, household, composition of adults, likely ages (from name analysis) and the number of 17 year olds – insufficient information for matching; (for details of current activity in the British social information industry, see Sleight 1991). The largest of the more detailed commercial databases is NDL's (National Demographics and Lifestyles) Lifestyle Database (Patron 1991); this has detailed records on only one in three households, a proportion which is probably near the maximum possible by voluntarily returned lifestyle questionnaires. It only holds 5 of the 24 pieces of information available on the census: age, sex, marital status, age of children and access to cars or vans; it merges occupation and economic activity status in such a way as to make it unmatchable with census information (Patron, personal communication, August 1992). Some Government managed registers (such as the National Health Service Central Register) are more comprehensive but contain relatively few variables and are maintained under conditions of strict security to ensure their confidentiality (Office of Population Censuses and Surveys 1992). In principle, investigators could attempt to compile a list themselves

via their own fieldwork, but this seems likely to be highly impractical for the sizes of areas usually identified (e.g., no less than 120,000 individuals in Britain).

One can concede immediately that the situation might differ in other countries, and that the situation could change in Britain. But the solution would be to group categories of census variables released on microdata to ensure that matches against a vulnerable population list could not be made. If dentists' professional association released a list of registered members containing their age, sex, and area of residence, one would simply need to ensure that dentists were not separately identifiable as an occupational group on the microdata sample that was released.

4.2. *Statistical inference*

An alternative approach which does not require the use of complete population information is statistical inference. In this approach it is recognized that population uniqueness is a characteristic of the population distribution of the key variables and the problem of making inference about this population characteristic given the sample microdata is a problem of statistical inference. An analogous inferential problem for continuous variables with measurement error is discussed by Paass (1988), Duncan and Lambert (1989) and Sullivan and Fuller (1989). For categorical variables Bethlehem et al. (1990) consider the estimation of the overall proportion of population uniques and Duncan and Lambert (1986) consider prediction of individual values. Here we are concerned with the prediction of population uniqueness for given individuals.

To take an informal example of how inference might proceed, consider again the example of the black female judge, aged 45, for which just one single exact match had been found in the microdata. Furthermore, make the (unlikely) supposition that prior information is also available on number of children, type of accommodation and number of rooms in the judge's dwelling. Suppose this information also matches exactly with the microdata record. It might be argued that this would be very unlikely were the microdata record to be for a different individual and hence that population uniqueness could be inferred with a reasonable degree of confidence. We take a more formal approach to statistical inference shortly.

4.3. *Figures in the public eye*

A third approach would be to argue that for certain conspicuous target individuals, any other individual with the same combination of census characteristics should be public knowledge. Thus, all local public figures with certain recognizable characteristics, such as a female veterinarian, a person from an unusual ethnic group aged 101 or a chief of police with nine children and a Ph.D., might expect to be "publicly" known in a local district. Similarly, certain occupational categories (politicians, actors and musicians) contain persons who are in the public eye and such persons with unusual characteristics might be expected to be publicly known at a national level.

As discussed by U.S. Department of Commerce (1978, p. 29), census offices might largely hope to protect against any confident inference of this kind by grouping categories in various ways. First, restricting the size of area identified helps to protect against the use of local knowledge. Second, recognizable groups, such as high profile occupational categories, could be combined with other less visible groups. Third, extreme values of

variables such as age, income or number of children can be "top-coded."

In summary, the only practical way in which an investigator might infer uniqueness with confidence appears to be the method described in Section 4.2. The other ways appear to be largely preventable in advance by census offices. We now examine this method in more detail.

The aim is to infer whether a given individual i has a combination of key variable values which is unique in the population. One approach would be Bayesian, following Duncan and Lambert (1986). An alternative frequentist approach, analogous to our discussion of identification risk in Section 2, would be to estimate the proportion of population unique individuals within some subpopulation S containing i. If the population values are assumed to be generated by a model, this proportion might be replaced by Pr(population unique$|S$): the probability that an individual is population unique given that event S applies to the individual. If this probability is high, subject to the error involved in estimation, the investigator could claim to have inferred population uniqueness for the individual. In the simplest case, S may be taken to be the whole population. If Pr(population unique) is not high then S might be taken to be the event that i is sample unique and Pr(population unique|sample unique) estimated (of course, if i is not sample unique then Pr(population unique|not sample unique) = 0). If this is not high then an even more restrictive S might be taken and so on.

To illustrate the estimation of Pr(population unique) from the microdata, we consider how inference might proceed under the Poisson-gamma model of Bethlehem et al. (1990). This model assumes that the frequency of occurrence of each combination of key variables has a Poisson distribution with a rate which varies between different combinations according to a gamma distribution with parameters α and β. Under this model

$$P = \Pr(\text{population unique}) = (1 + N\beta)^{-(1+\alpha)} \tag{4}$$

where N is the population size (see Appendix). To illustrate how α and β and hence $P = \Pr(\text{population unique})$ can be estimated we have used the Italian census microdata for Tuscany described in Marsh et al. (1991). We selected a random sample of 10,000 individuals from 3.5 million cases to represent an artificial microdata sample and estimated α and β using a procedure described in the Appendix. Eight key variables were selected: area, age, sex, marital status, housing tenure, occupational group, employment status and household structure. Three levels of detail were considered for both area (1, 2 or 9 areas identified) and age (1, 5 or 10 year bands), making nine sets of key variables altogether. The estimates of α and β for each set of key variables are shown in Table 1 along with the estimated values $\hat{P}$ of P obtained by substituting these estimates and $N = 3.5 \times 10^6$, the original population size, into formula (4), together with the "true" values of P estimated (with negligible error) from the 3.5 million cases.

The estimates $\hat{P}$ are similar to the true values although a systematic bias is evident. Plotting both estimates and true values against $p = \Pr(\text{sample unique})$, for example, shows clearly that as p increases beyond 15%, say, there is increasing underestimation of P. Thus, for $p = 54.1\%$, the approximate 95% confidence interval of $0.33 \pm 2 \times 0.007 = (0.32\%, 0.35\%)$ severely underestimates the true value of 0.98%. In the Appendix we consider a number of possible reasons for this bias and conclude that the Poisson-gamma model must be

Table 1. Estimates of $P = Pr(\text{population unique})$ from Poisson-gamma model

Areas identified	Age band	$\hat{\alpha}$ (10^{-3})	$\hat{\beta}$ (10^{-4})	P (%)	$\hat{P}$ (%)	S.E. (%)	Pr(sample unique) (%)
1	1 year	5.15	3.35	0.14	0.08	0.002	22.8
	5 years	9.34	9.10	0.01	0.03	0.001	9.7
	10 years	12.45	13.04	0.00	0.02	0.001	6.9
2	1 year	3.92	2.20	0.36	0.13	0.003	31.1
	5 years	7.56	5.63	0.06	0.05	0.001	14.9
	10 years	9.10	8.92	0.03	0.03	0.001	9.9
9	1 year	2.26	0.85	0.98	0.33	0.007	54.1
	5 years	3.63	2.60	0.26	0.11	0.002	27.6
	10 years	4.99	3.61	0.14	0.08	0.002	21.5

misspecified in some way. It is, however, beyond the scope of this paper to attempt to find a better alternative. The only conclusion we wish to draw from Table 1 is that, because the estimated values of P are at least in the right ball park and because improved estimation might be possible with alternative models, it should be assumed feasible, without further evidence to the contrary, that an investigator could infer the value of P from the microdata.

The important point to note is that, for values of P as low as in Table 1, the investigator will be unable to use such values to infer with any confidence that a link is correct for any given target individual.

Now, as noted earlier, if the estimated value of P turns out to be small, the investigator could try instead to estimate the conditional probability of population uniqueness given sample uniqueness. Proportions of individuals that are sample unique are shown in the last column of Table 1. For such individuals, the probability of population uniqueness is shown in the Appendix to be raised to

$$\text{Pr(population unique|sample unique)} = \left(\frac{1+\theta n/N}{1+\theta}\right)^{1+\alpha} \tag{5}$$

where $\theta = N\beta$ and n is the size of the microdata sample. As we would expect this is the same as P in (4) if $n = 0$ but is larger than P if $n > 0$ (it is assumed in the Poisson-gamma model that $\alpha > 0$, $\beta > 0$). For example, for the values $\alpha = 2.26 \times 10^{-3}$, $\theta = 297.5$ above, we have $P = 0.33\%$, and if $n/N = 0.02$, we have Pr(population unique| sample unique) = 2.33% which is greater. However, this probability is still far too low to infer uniqueness with confidence.

The increase in the ability to infer population uniqueness which an investigator gains by restricting attention to sample uniques is a function of the microdata sampling fraction for given α and θ. In Figure 1 the value of Pr(population unique|sample unique) is plotted as a function of n/N for fixed values of $\alpha = 2.26 \times 10^{-3}$ and $\theta = 297.5$. The function increases from $P = 0.33\%$ when $n/N = 0$ up to 1 when $n/N = 1$. In general, the function is non-linear but the value of α here is so small that the function is visually indistinguishable from a straight line. For any values of α, $\beta > 0$, the function is always convex and this implies that

$$\text{Pr(population unique|sample unique)} \leqslant P + (1-P)n/N.$$

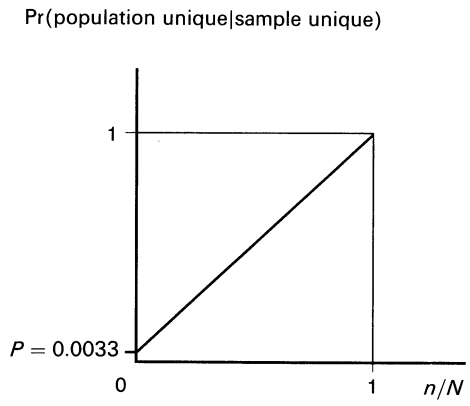

Fig. 1. Probability of population uniqueness given sample uniqueness as a function of the sampling fraction

Hence, whenever the sampling fraction n/N is small, only a small improvement in the investigator's confidence in population uniqueness can be made by restricting attention only to sample uniques.

If the estimated probability in (5) is small, then the investigator could restrict the subpopulation S still further and, in the limit, attempt to specify and estimate a model relating population uniqueness as a binary response to all the key variables as explanatory variables. This would appear to be difficult in the absence of population microdata but if it were possible then the investigator could use the fitted model to predict the probability of population uniqueness for any target individual with given key variable values and, in principle, to select that individual for which this predicted probability was maximum.

5. Estimation of Identification Risk

Census authorities considering the release of census microdata would like to estimate the identification risk for alternative forms of release, e.g., for different choices of categories of variables. Estimation of this risk will necessarily involve judgements about the possible prior information available to potential investigators and the possible scenarios of attempts at disclosure. One direct approach to estimation involves the use of empirical experiments such as in Blien et al. (1992). Such experiments may not only be very expensive, however, but may also be impossible, for example, because a decision about release is required before the data are collected or because practical or legal reasons prevent the linking of identifiers. There is, therefore, a need for simpler approaches to estimation, and we now indicate one approximate approach, first outlined in Marsh et al. (1991).

Let us suppose that the key variables and the subpopulation S, from which the target individual is randomly selected, are specified. Given our constraint on false positives, we may approximate the identification risk by the proportion of individuals in S who would be correctly identified by a reliable identification rule. Assuming a rule based on exact matching, it is necessary for correct identification to take place that the following two events occur.

A: the individual appears in the microdata,
B: the key variable values for the individual are recorded identically in the prior information and the microdata.

Furthermore, as argued in Section 3, in order to verify that the link is correct it is necessary to verify that the individual is population unique and a necessary condition to be able to achieve this reliably is that the individual is population unique, i.e.,

C: the combination of key variable values for the individual is unique in the population.

These events together with the further event:

D: the investigator is able to verify with high probability that the individual is correctly linked.

All of these events may be taken as approximately necessary and sufficient for identification and we may write:

$$\text{Identification risk for individual } i \doteq \Pr(A \text{ and } B \text{ and } C \text{ and } D)$$

or in terms of a succession of conditional probabilities:

$$\text{Identification risk} = \Pr(A)\Pr(B|A) \times \Pr(C|A,B)\Pr(D|A,B,C). \quad (6)$$

The probabilities are interpreted, as before, as the proportion of individuals in the subpopulation S for whom the respective events apply. We suggest that rough estimates of the first three components of risk $\Pr(A)$, $\Pr(B)$ and $\Pr(C|A,B)$ will often be obtainable in practice and we discuss this in the following three subsections. The fourth component is more difficult to assess and requires some qualitative judgement in the light of the discussion of Section 4. There seems room, however, for some quantification. For example, model-based methods as discussed in Section 4 might be used to estimate from the microdata the probability that an individual is population unique for individuals in S who obey conditions A, B and C. The proportion of such individuals for which this estimated probability is above 90%, say, might be taken as an estimate of the probability of verifying population uniqueness given A, B and C.

In any case, $\Pr(A, B, C)$ provides an approximate upper bound for the identification risk. For related discussion, see Greenberg and Voshell (1990, p. 1) who "regard the number of population uniques present on the microdata file as one of the components of a measure of disclosure risk."

5.1. *Event A: presence in the microdata*

In the simplest case, the microdata sample is selected by a randomised equal probability sampling procedure so that the first component of the identification risk in (6) is

$$\Pr(A) = 1/k$$

where $1/k$ is the sampling fraction. For example, $k = 50$ in the proposed British individual file. In practice, census nonresponse and population change may tend to reduce $\Pr(A)$ somewhat. For example, Blien et al. (1992) only found 53 individuals or 0.66% of the 7,983 individuals in their handbook to be present in the microcensus sample even though the sampling fraction in the microcensus was 1%.

We may usually expect $\Pr(A)$ to be unaffected by the choice of subpopulation S since the randomised selection of the sample will prevent "response knowledge" (Bethlehem et al. 1990). However, sometimes there will be variations in inclusion probabilities which might in some circumstances be useable by a malicious investigator to increase the conditional probability of A given choice of S.

For example, for the 1980 U.S. census or the 1986 Canadian census, the samples are selected only from those households which complete the long form of the census schedule, which is only administered to around one household in five. Thus, in the (unlikely) event that the investigator was able to ascertain that a target individual had filled in a long form, the prior inclusion probability would be increased five-fold. Also, for some microdata the sampling fraction varies between geographical regions, for example, in the 1986 Canadian census the fraction varied between 1/20 and 1/100 (Denis 1989).

5.2. *Event B: identical recording of key variables in both datasets*

There are three possible reasons why the vectors $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ may differ, even though the investigator may have attempted to use

key variables which have common definitions in the two datasets.

First, variables which appear in theory to have the same definition may in practice be operationalised in different ways: many different definitions of social class exist, for example, (Marsh 1986), and even less contentious variables such as long-term illness or access to a car are often measured in very different ways.

Second, there may be a time mismatch between the measurement of the variables on the two datasets. Almost all the variables regularly collected on censuses are subject to change. The only common time-invariant census variables are sex, date of birth, and place of birth, while some, such as labour force participation status, can change frequently. While in theory the owner of a large database could freeze the information about the known individuals at the date of the census in order to facilitate matching, in practice these multi-purpose databases are constantly updated, often with information whose precise dating is either not clear or not recorded. Two years commonly elapse between the date of the census and the earliest date by which a user might expect to receive micro-data; this time lag not only reduces the probability of a perfect match but may also be expected to reduce the probability of attempts at disclosure (U.S. Department of Commerce 1978, p. 27; Bethlehem et al. 1990).

Finally, the variables in either dataset may be subject to measurement error (including respondent error, coding error, typing error and so on). The problems of measurement error in census variables have long been recognized (Hansen, Hurwitz, and Bershad 1961). Census post-enumeration surveys in Britain and the United States illustrate the problems that can arise with seemingly straightforward variables such as number of rooms, housing tenure or access to a car (Britton and Birch 1985; U.S. Department of Commerce 1986). Further evidence of response variation in census-type variables such as ethnicity, employment status, and housing characteristics is given by Turner and Martin (1984, chs. 4, 5 and 6) and Martin, DeMaio and Campanelli (1988).

While it seems reasonable to suppose that event B will usually be independent of event A so that $\Pr(B|A) = \Pr(B)$, it seems less plausible that $\Pr(B)$ will be independent of the choice of subpopulation S from which the individual is randomly selected. Fortunately for disclosure control, it seems likely that $\Pr(B)$ will often actually be reduced for those subpopulations with rare characteristics which a malicious investigator is most likely to choose to focus on. For it seems to be a common property of measurement error that rare categories of variables often display a higher error rate than common categories. For example, comparing linked 1971 and 1981 British census records suggests that the error rates for country of birth are lowest for England and highest for the less common countries of birth (Office Population Censuses and Surveys 1988). Similarly, the 1980 U.S. post-enumeration survey showed higher inconsistency rates with rare tenures such as living in accommodation rent-free than owning one's own accommodation (U.S. Department of Commerce 1986, p. 77). This property tends to reduce the probability of condition B being achieved for just those individuals with rare census characteristics whom a mischievous investigator might try to track down. Rare categories also arise by crossing univariate categories, for example, female dentists in a given town. Such multivariate categories may also tend to have higher measurement errors than the corresponding univariate categories

because of cumulation of errors. A simple attempt to illustrate the possible magnitude of cumulative measurement error on several variables was made in Marsh et al. (1991). On the basis of error rates estimated from post-enumeration surveys for five key variables: household size, tenure, number of cars and vans, economic activity and social class, it was estimated that the probability that all the variables would be correctly classified in both datasets would be around 0.64, even using the same variable definitions and taking measurements at the same point in time.

Blien et al. (1992) found in their study linking the German microcensus with a handbook of scientists and scholars that out of 53 common individuals, 43 had incompatible values on at least one of the ten key variables so that the implied estimate of $\Pr(B|A)$ is 0.19.

5.3. *Event C: population uniqueness*

Just as for event B, it seems reasonable to suppose that event C is independent of event A so that $\Pr(C|A,B) = \Pr(C|B)$. Also, given the tendency discussed in the previous subsection for rare characteristics to have high error rates, we may expect events B and C to be negatively associated so that $\Pr(C)$ can act as an approximate upper bound for $\Pr(C|B)$.

Let us first consider the case when the subpopulation S, from which the target individual is randomly selected, consists of the whole population. In this case, $\Pr(C)$ is equal to the proportion of individuals in the population with unique combinations of values on the key variables. This assumption may be a reasonable first approximation for the first type of potential investigator who wishes to use record linkage to add information to an arbitrary record from an existing large database.

The value of $\Pr(C)$ will depend on the number and detail of the key variables and on the size of the population. Schlörer (1975) carried out an experiment using random subsets of 10 census-type key variables (with an average of 8.5 categories) for a population of 31,465 individuals and found that, using 4, 5, 6, 7, or 8 key variables, the proportions of unique combinations were 1%, 11%, 32%, 56%, or 76% respectively. In a population of 23,485 households composed of a father, mother and two children, Bethlehem et al. (1990) found the proportion of households with unique combinations of 6 key variables, consisting of ages of the father and mother and ages and sexes of the two children (all ages in years), to be 68%. For a larger population of 87,959 records from the 1980 U.S. Census, Greenberg and Voshell (1990) found, using 6, 10, or 15 key variables, proportions of unique combinations of 0.4%, 9%, and 35% respectively. Which of the above proportions of uniqueness is the most realistic estimate of $\Pr(C)$ for a specific census microdata file clearly depends on the number of key variables and on the size of the subpopulation which can be identified by further key variable information. Marsh et al. (1991) conducted an experiment with Italian census microdata using eight key variables which were judged to err on the cautious side on the basis of an investigation of the availability of variables in commercial databases. For a population of 3.5 million individuals with nine areas identified to a minimum size of 200,000 and with age recorded in one year bands the estimated proportion of uniques in the population was about 1%.

Let us now turn to the second type of potential investigator, who wishes to achieve disclosure for its own sake and who, it is at least conceivable, might

select a target individual from some subpopulation S, defined by categories of key variables which are perceived to be unusual. Marsh, Dale, and Skinner (1993) carried out an experiment on British census data with six key variables and with different choices of S consisting of the different categories of each of the key variables, one at a time. They found that the proportion of unique cases was indeed greater for unusual categories and furthermore that the proportion could be predicted quite reliably from the relative size of the category. The implication for disclosure control was that the fineness of the category sizes should be restricted.

When evaluating specific proposals for the release of census microdata, we suggest that the census agency conduct a similar study to those above, assessing the sensitivity of the estimated value of $\Pr(C)$ to alternative plausible types of key variable information, to alternative amounts of detail in the microdata and to alternative possible subpopulations S. See Dalenius (1986) for computational methods of determining population uniqueness.

6. The Alternative Scenario

In Sections 3–5 we have investigated the identification risk for scenario (a) of Section 2 in which an investigator, who holds prior data on a target individual, seeks a matching record from the microdata. Alternatively, the investigator might employ scenario (b) and first select a microdata record before seeking a matching individual from the general population. Ironically, the only obvious rationale for doing this would be a demonstrative breach of confidentiality.

Let us suppose then that an investigator selects a record from the microdata and sets out to identify the person in the population to whom this record belongs. One approach would be to search a database containing prior information on a number of individuals in the population. In this case, the probability of success can be assessed in an identical way to that discussed above, but reversing the roles of the microdata sample and the database. Following the approximate approach in Section 5, the only element in equation (6) that would change would be $\Pr(A)$, which is now given as the proportion of the population in the database rather than the proportion appearing in the microdata sample.

However, unlike the microdata sample which was selected on a random basis, it would now be possible to select an individual from the microdata deliberately in order to increase $\Pr(A)$ – the probability of the microdata respondent appearing in the database. For example, if the investigator had a reasonably comprehensive list of car owners, by selecting a car owner from the microdata, $\Pr(A)$ would approach 1. But, as noted in Section 4 under (1), databases which have reasonably comprehensive coverage of subgroups defined by census variables, such as government registers, often contain few key variables, so that an increased value of $\Pr(A)$ is compensated for by small values of $\Pr(C|A,B)$. An alternative way to seek a match would be by advertising for a person with particular characteristics to come forward and identify himself or herself. However, it would seem essential for the person concerned to reveal his or her characteristics voluntarily for this approach to work. Disclosure could not be achieved without the person's consent. This is not strictly a breach of confidentiality on the part of the census agency, and we therefore can exclude it from the current assessment of risk.

7. Discussion

Disclosure control may be achieved by

varying the factors, such as the sampling fraction or the definition of categories of variables, which affect the identification risk. Given the definition of risk in equation (3), the effects of these factors need to be assessed both in terms of their effect on the proportion of successful identifications for a given identification rule and also in terms of their effect on the class of possible rules for which a reliable and verifiable identification is feasible. The latter may involve checking for the existence of: (i) lists with complete coverage of individuals in subgroups of the population defined by census variables and (ii) categories of variables such as occupation, which might identify figures in the public eye.

The estimation of identification risk depends crucially on what variables are considered to be key variables. This implies the need for continued surveillance of external databases which might contain census variables. It might be argued that such surveillance is impractical. Thus, McGuckin and Nguyen (1988, p. 196) state that "there is no easy way to know exactly what information is available to the public nor is there any easy way to evaluate its quality or how well it can duplicate Census data. Moreover, it is impossible for an agency that wants to release a public use microdata file to keep track of new outside files and changes to existing ones." Whilst we recognize this difficulty, it is not clear that it is always an impossibility. In Britain, all personal data which are stored on a computer are subject to the provisions of the Data Protection Act 1984. The Act gives individuals rights to find out and if necessary challenge the information held on them. Those who hold such personal data are required to register as data users and to state the purposes for which they hold the data, and this register is open to the public (Data Protection Registry 1989a). As a result, the most important holders of data, and the information contained in their databases, are relatively easily discovered.

All European countries which have ratified the Council of Europe Convention on Data Protection have agreed to abide by a set of common principles governing such personal data. While the Convention itself does not cover the process of registering data users (Data Protection Registry 1989b), many countries in Europe have set up registration procedures similar to those operating in Britain, which would facilitate discovery of the major users of personal data. While it is true that the register might not be comprehensive – it might not be completely up to date, new information might have been added to a database, some private sector companies might not be on the register because in some countries they are not obliged to register – nonetheless the existence of such registers and of knowledgeable staff to run them means that in practice it is most unlikely that any major data gathering exercise could escape notice.

A specific application of our general approach to proposals for the release of census microdata in Great Britain is described by Marsh et al. (1991), who conclude that it would be feasible to produce microdata which are both valuable to potential users and for which the risk of disclosure is "very small."

Our approach to disclosure control has assumed methods which preserve the integrity of the data in the terminology of Section 1. The alternative approach to disclosure control is by contamination methods, which cannot only reduce the risk of identification, but also break the automatic link between identification and disclosure described in Section 2. Thus, even if an investigator could correctly link an individual to a microdata record, it would still be difficult to disclose new

information with confidence if there had been substantial contamination.

The most important disadvantage of this approach is that it can grossly distort statistical analyses when substantial contamination is applied. Contamination is usually controlled so that certain predetermined characteristics of the original data, such as means, variances and covariances (Kim 1986) and univariate distribution functions (Sullivan and Fuller 1989) are preserved. This may create no difficulties for microdata released for specific applications, but the usual purpose of census microdata is to provide a broad and flexible data resource for uses which cannot be determined in advance. Large numbers of tabulations are produced to meet established demand, but microdata might be used, for example, for the analysis of subsets of data on minority groups which become of topical policy interest or to fit binary response regression models which might be selected only in the light of exploratory data analysis of the microdata. While predetermined characteristics such as means and variances may not be affected by contamination, uncontrolled characteristics, such as binary regression relationships, may be.

Appendix

A1. Poisson-gamma Model

Following Bethlehem et al. (1990) let K be the number of possible combinations of key variable values in the population. Let F_i be the number of individuals in the population with combination i, $i = 1, \ldots, K$ and let f_i be the corresponding number in the microdata. We assume that combination i occurs with probability π_i where the values of $\pi_1, \ldots, \pi_K$ are generated by a gamma distribution with parameters α and β, such that $K\alpha\beta = 1$. We assume that F_i, f_i and $F_i - f_i$ are Poisson distributed conditional on π_i

$$F_i|\pi_i \sim \text{Poisson}\,(N\pi_i),$$

$$f_i|\pi_i \sim \text{Poisson}\,(n\pi_i),$$

$$(F_i - f_i)\pi_i \sim \text{Poisson}\,[(N-n)\pi_i]$$

and that f_i and $F_i - f_i$ are independent given π_i. It follows, by integrating out π_i, that

$$\Pr(F_i = 1) = N\alpha\beta(1 + N\beta)^{-(1+\alpha)}.$$

Note that this expression differs from equation (6.4) of Bethlehem et al. (1990), since we use the usual gamma parameterization (e.g., Johnson and Kotz 1969, p. 125), whereas the parameters α_B and β_B, say, used by Bethlehem et al. (1990), appear to be $\alpha_B = \alpha/N$, $\beta_B = \beta N$. The expected number of individuals in the population for which $F_i = 1$ is $K\Pr(F_i = 1)$ and the probability that an individual, selected at random from the population, has a unique combination is

$$\begin{aligned} P &= \Pr(\text{population unique}) \\ &= N^{-1}K\Pr(F_i = 1) \\ &= (1 + N\beta)^{-(1+\alpha)} \end{aligned}$$

which is equation (4). Similarly, the probability that an inividual randomly selected from the microdata sample has a combination which is unique in the sample is

$$\begin{aligned} \Pr(\text{sample unique}) &= n^{-1}K\Pr(f_i = 1) \\ &= (1 + n\beta)^{-(1+\alpha)}. \end{aligned} \tag{A1}$$

Using the fact that

$$\begin{aligned} &\Pr(F_i = 1, f_i = 1|\pi_i) \\ &\quad = \Pr(F_i - f_i = 0|\pi_i)\Pr(f_i = 1|\pi_i) \end{aligned}$$

we obtain similarly

$$\Pr(F_i = 1,\ f_i = 1) = n\alpha\beta(1 + N\beta)^{-(1+\alpha)}$$

and so the probability that an individual randomly selected from the sample has a

combination which is both unique in the sample and in the population is

$$\begin{aligned}&\Pr(\text{sample and population unique})\\&= n^{-1}K\Pr(F_i = 1,\ f_i = 1)\\&= (1 + N\beta)^{-(1+\alpha)}. \qquad \text{(A2)}\end{aligned}$$

Equation (5) then follows by dividing (A2) by (A1).

A2. Estimation

Estimates $(\hat{\alpha}, \hat{\beta})$ of (α, β) were obtained by equating p, the observed proportion of sample uniques, to formula (A1), substituting $\beta = 1/(K\alpha)$, and solving for α using Newton's iterative method. The estimate $\hat{P}$ of P was then obtained by substituting $(\hat{\alpha}, \hat{\beta})$ for (α, β) in (4). The standard error of $\hat{P}$ was estimated by the usual δ-method as $c_p v(p)^{1/2}$, where straightforward algebra gives

$$v(p) = p(1-p)/n, \quad c_p = \hat{P}a(N)/[pa(n)],$$

$$a(n) = \log(1 + n\hat{\beta})/(K\hat{\beta}^2) - [1 + 1/(K\hat{\beta})]n/(1 + n\hat{\beta}).$$

A3. Remarks on Lack of Fit of Model

On observing the systematic bias of $\hat{P}$ as an estimator of P in Table 1, it might first be asked whether alternative estimation procedures, such as method of moments or maximum likelihood, as considered by Bethlehem et al. (1990), might have given different results. However, if the model and assumptions are correct, then all these estimators should be consistent and only differ in their implied confidence interval widths. Given our narrow confidence intervals, it is necessary to question our model and assumptions, rather than our estimation procedure.

One assumption that is very questionable is that we know K. The values we use for K exclude certain combinations of key variables, such as married two-year olds, which we judge logically impossible. But such judgements are inevitably somewhat subjective. In fact, it turns out that $\hat{P}$ is extremely insensitive to the choice of K. For example, for Provinces and 1-year age bands, the effect of doubling K is to change $\hat{P}$ from 0.332 to 0.334 ($\hat{\alpha}$ is roughly halved and $\hat{\beta}$ is hardly affected). This insensitivity may be seen alternatively by allowing for K to be estimated with an error which has variance v_K and is uncorrelated with p. In this case, the estimated standard error of $\hat{P}$ becomes

$$[c_p^2 v(p) + c_K^2 v_K]^{1/2}$$

where $c_K = \hat{P}[n\log\hat{P}/(1 + n\hat{\beta}) - N\log p/(1 + n\hat{\beta})]/(a(n)K^2\hat{\beta})$.

Under the very conservative assumption that $v_K = K^2$, the standard error of the value of $\hat{P}$, above, only increases from 0.0066 to 0.0074. The reason for the insensitivity of $\hat{P}$ to K is that if α is very small, as in our case, then $p \doteq (1 + n\hat{\beta})^{-1}$ so that $\hat{\beta} \doteq (p^{-1} - 1)/n$ and

$$\hat{P} \doteq [1 + N(p^{-1} - 1)/n]^{-1}$$

which does not depend on K. For example, this formula gives $\hat{P} = 0.335$ compared to the value 0.332 in the case above.

It seems, therefore, that the Poisson-gamma model itself must be questioned. One theoretically unattractive feature of the model is that the probabilities π_i, which lie in the interval [0, 1], are modelled by a gamma distribution, defined on the interval $[0, \infty]$. A more attractive model in this respect would be to assume that N is fixed, $F_1, \ldots, F_K$ are multinomial given $\pi_1, \ldots, \pi_K$ with parameters $(N, \pi_1, \ldots, \pi_K)$ and the marginal distribution of each π_i is beta with parameters a and b. However, if $E(\pi_i) = 1/K \to 0$ and $K \to \infty$ and if the coefficient of variation of the π_i converges

to a finite non-zero limit as $K \to \infty$, then a also converges to a finite non-zero limit and $b \to \infty$ as $K \to \infty$. But, in this case, the beta distribution converges to a gamma distribution with parameters $\alpha = a$ and $\beta = b^{-1}$ and the formulae for $\Pr(F_i = 1)$ and $\Pr(F_i = 1 | f_i = 1)$ become identical to those for the gamma distribution under this reparameterization. In other words, if we were to fit this beta-multinomial model to the microdata and estimate P as a function of the estimated a and b, we would expect almost identical answers to the Poisson-gamma model. Hence, it appears that it is some other aspect of the specification of the Poisson-gamma model than that the gamma distribution is not restricted to $[0, 1]$, that causes the bias in $\hat{P}$. As a referee notes, one possibility would be to extend the Poisson-gamma model to a mixture of such distributions.

8. References

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

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

Britton, M. and Birch, F. (1985). 1981 Census Post-Enumeration Survey. London: Her Majesty's Stationery Office.

Butz, W.P. (1985). Data Confidentiality and Public Perceptions: the Case of the European Censuses. Paper presented to the American Statistical Association, Las Vegas, August 5–8.

Courtland, S. (1985). Census Confidentiality: Then and Now. Government Information Quarterly, 2, 407–418.

Cox, L.D., McDonald, S., and Nelson, D. (1986). Confidentiality Issues at the United States Bureau of the Census. Journal of Official Statistics, 2, 135–160.

Cox, L.H. and Sande, G. (1979). Techniques for Preserving Statistical Confidentiality. Bulletin of the International Statistical Institute, 42.3, 499–512.

Dalenius, T. (1977a). Privacy Transformations for Statistical Information Systems. Journal of Statistical Planning and Inference, 1, 73–86.

Dalenius, T. (1977b). Towards a Methodology for Statistical Disclosure Control. Statistisk tidskrift, 5, 429–444.

Dalenius, T. (1986). Finding a Needle in a Haystack or Identifying Anonymous Census Records. Journal of Official Statistics, 2, 329–336.

Data Protection Registry (1989a). Data Protection Act 1984: Introduction to the Act, Guideline No. 1: Cheshire, U.K.

Data Protection Registry (1989b). Data Protection Act 1984: The Data Protection Principles, Guideline No. 4: Cheshire, U.K.

Denis, J. (1989). Fichier de Microdonnées à Grande Diffusion des Ménages et des Logements du Recensement de 1986: Méthodologie et Recommendations. Statistics Canada, Ottawa.

Duncan, G. and Lambert, D. (1986). Disclosure-Limited Data Dissemination. Journal of the American Statistical Association, 81, 10–28.

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

Fellegi, I. (1972). On the Question of Statistical Confidentiality. Journal of the American Statistical Association, 67, 7–18.

Gates, G.W. (1988). Census Bureau Microdata: Providing Useful Research Data while Protecting the Anonymity of Respondents. Proceedings of the Social

Statistics Section, American Statistical Association, 235–240.

Greenberg, B. and Voshell, L. (1990). The Geographic Component of Disclosure Risk for Microdata. SRD/RR-90/13. U.S. Bureau of the Census, Washington, D.C.

Hansen, M.H., Hurwitz, W.N., and Bershad, M. (1961). Measurement Errors in Censuses and Surveys. Bulletin of the International Statistical Institute, 38, 359–374.

Johnson, J.L. and Kotz, S. (1969). Discrete Distributions. New York: John Wiley.

Kim, J. (1986). A Method of Limiting Disclosure in Microdata Based on Random Noise and Transformation. Proceedings of the Survey Research Methods Section, American Statistical Association, 370–374.

Marsh, C. (1986). Occupation and Social Class. In R. Burgess (Ed) Key Variables in Social Research, 2nd Ed., London: Routledge and Kegan Paul, 123–152.

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

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991). The Case for Samples of Anonymized Records from the 1991 Census. Journal of the Royal Statistical Society, Ser. A, 154, 305–340.

Martin, E., DeMaio, T.J., and Campanelli, P.C. (1988). Context Effects for Census Measures of Race and Hispanic Origin. Paper presented at Annual Meetings of American Statistical Association, New Orleans, August 1988.

McGuckin, R.H. and Nguyen, S.V. (1988). Use of "Surrogate Files" to Conduct Economic Studies with Longitudinal Microdata. Proceedings of the Bureau of the Census 4th Annual Research Conference, Bureau of the Census, Washington, D.C. 193–211.

Office of Population Censuses and Surveys (1988). Census 1971–81: The Longitudinal Study. London: Her Majesty's Stationery Office.

Office of Population Censuses and Surveys (1992). Statement of Policies on Confidentiality and Security of Personal Data, Mimeo, London.

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

Patron, M. (1991). Customised Modelling of Census Data. In Sleight (1991), 130–133.

Schlörer, J. (1975). Identification and Retrieval of Personal Records from a Statistical Data Bank. Methods of Information in Medicine, 14, 7–13.

Skinner, C.J. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. Statistica Neerlandica, 46, 21–32.

Sleight, P. (1991). Using the New Census Data to Improve Targeting. Proceedings from Henry Stewart Conference Studies, 2/3 Cornwall Terrace, Regent's Park, London, December 6.

Spruill, N.L. (1983). The Confidentiality and Analytic Usefulness of Masked Business Microdata. Proceedings of the Survey Research Methods Section, American Statistical Association, 602–607.

Strudler, M., Oh, H.L., and Scheuren, F. (1986). Protection of Taxpayer Confidentiality with Respect to the Tax Model. Proceedings of the Survey Research Methods Section, American Statistical Association, 375–381.

Sullivan, G.R. and Fuller, W.A. (1989). The Use of Measurement Error to Avoid

Disclosure. Proceedings of the Survey Research Methods Section, American Statistical Association, 802–807.

Turner, C.F. and Martin, E., Eds. (1984). Surveying Subjective Phenomena, Vol. 2. New York: Russell Sage Foundation.

U.S. Department of Commerce (1978). Report on Statistical Disclosure and Disclosure Avoidance Techniques. Statistical Policy Working Paper 2, Washington, D.C.

U.S. Department of Commerce (1986). Content Reinterview Study: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview. 1980 Census of Population and Housing, Bureau of the Census, PHC890-E2.

Willenborg, L.C.R.J., Mokken, R.J., and Pannekoek, J. (1990). Microdata and Disclosure Risks. Proceedings of the Bureau of the Census 1990 Annual Research Conference, Bureau of the Census, Washington, D.C., 167–180.

Received June 1990
Revised August 1993