1352.0.55.060

Australian Bureau of Statistics

Research Paper

Assessment of Risk for Unit Record File Disclosure

Spontaneous Recognition and Population Modelling

www.abs.gov.au

New Issue

Research Paper

Assessment of Risk for Unit Record File Disclosure

Spontaneous Recognition and Population Modelling

Paul Schubert

Statistical Services Branch

Methodology Advisory Committee

21 November 2003, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 5 JUL 2004

ABS Catalogue no. 1352.0.55.060
ISBN 0 642 48164 4

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Paul Schubert, Statistical Services Branch on Canberra (02) 6252 6591 or email <paul.schubert@abs.gov.au>.

ASSESSMENT OF RISK FOR UNIT RECORD FILE DISCLOSURE – SPONTANEOUS RECOGNITION AND POPULATION MODELLING

Paul Schubert
Statistical Services Branch

EXECUTIVE SUMMARY

The ABS makes unidentifiable microdata from its surveys available to users in the form of Confidentialised Unit Record Files (CURFs). To confidentialise these microdata, we aim to protect the data against two specific scenarios: spontaneous recognition and matching to lists.

Spontaneous recognition of an individual occurs when a user, while looking at information on a microdata file, recognises a particular record as possibly corresponding to a particular person that they know of. In order to assess the risk of spontaneous recognition, we consider data items which may need to be collapsed or masked in some way on the basis of how many individuals we estimate there to be in the population in a particular small-dimensional cross-classification.

A method we are considering for assessing the risk of identification of individuals from list matching is what we refer to as population modelling – or, more fully, modelling population probabilities based on sample survey data. In particular, we require the method to be able to identify whether combinations of data items that are observed for individuals in the sample are likely to be unique in the population. Because of the unusual application of the modelling, as well as it being the least developed aspect of our work so far, we would appreciate comments and guidance from MAC on Section 2 of the paper in particular. The paper concludes with a discussion of the applications of population modelling to confidentiality issues.

Note that this paper describes thoughts on assessing risk via spontaneous recognition and list matching for unit record file disclosure as at November 2003. It does not necessarily describe ABS practice or procedures at that time, or in the future.

DISCUSSION POINTS FOR MAC

Specific discussion points for MAC are highlighted throughout the paper but are listed here for convenience.

- Does MAC agree that a high level of detail of common knowledge variables does not need to be considered for spontaneous recognition?
- Does MAC agree with our characterisation of spontaneous recognition?
- Does MAC have any comments about using the expected proportion of population uniques from the sample as the measure to assess, rather than the probability of being a unit being population unique given that it is a sample unique as some other international agencies do?
- Does MAC have any comments on model selection and fitting for graphical models to be used to predict population uniques?
- Does MAC have any comments on our proposed validation approach?
- Does MAC have any ideas for best measures of information loss?

CONTENTS

ASSESSMENT OF RISK FOR UNIT RECORD FILE DISCLOSURE – SPONTANEOUS RECOGNITION AND POPULATION MODELLING

Paul Schubert
Statistical Services Branch

INTRODUCTION

Like other national statistical agencies, the ABS is faced with a tension in the area of public availability of its microdata. On the one hand, a prime reason for its existence is to collect high quality data to inform national policy and enable statistical research. This goal necessitates dissemination of microdata as well as summary data. On the other hand, the ABS is charged with protecting the confidentiality of its data providers, its survey respondents. This is explicit in its legislation – release of microdata is required to be 'in a manner unlikely to enable identification of individuals or organisations' – but is also important because public trust and perceptions of that trust are important contributors to data quality and response rates. Note that the ABS legislation focuses on identification – even if no additional information is disclosed.

There are four key ways by which the ABS aims to manage the risk of identification:

i. by protecting the data themselves e.g. collapsing categories, masking particular records by changing unusual characteristics, deleting records, withholding sensitive data items, restricting the detail of identifying items such as geographic location;

ii. by protections built in to the way the microdata may be accessed e.g. security and output restrictions applied to the Remote Access Data Laboratory (RADL), and data laboratories on-site at the ABS (ABSDLs);

iii. by influencing user behaviour through the use of signed undertakings and training material (e.g. Responsible use of Confidentialised Unit Record Files manual); and

iv. by regular audits of use of microdata, and sanctions such as removal of access to microdata or legal action.

There are three different modes of access to unidentifiable microdata offered by the ABS. These are via CD-ROM, via the RADL and via ABSDLs. The combination or mix of the four layers of protection differs with each mode of access.

A June 2002 MAC paper – *Protecting Confidentiality of Data Accessed from the Remote Data Centre* – focused on the protections associated with the RADL, part of the second of the above points. This paper focuses on the first of the four points: protections applied to the data themselves.

There are two situations that the ABS aims to guard against. The first is spontaneous recognition. This is the situation where a researcher or user is browsing through the microdata and happens to notice some unusual characteristics – generally a combination of a small number of data items – which remind her/him of someone that they know. Looking further at other data items can confirm the identity. The ABS aims to protect against spontaneous recognition for all three modes of access.

The second situation to guard against is the risk of identification through matching to administrative lists. Administrative lists often contain the same or similar demographic data items (age, sex, geographic location, etc.) as the ABS microdata. The two datasets could be matched via these variables and, in cases where the matches are one-to-one, the combined set of data items can pose an additional disclosure risk. CD-ROM is the only mode of access to microdata that is at high risk from matching to administrative lists; the environments of the other two modes of access make it very difficult if not impossible to perform such a match without being detected. The risk of identification, and hence the amount of confidentialising action required, is generally much greater from list matching than it is from spontaneous recognition. Thus the main, although not exclusive, focus for assessing basic CURFs for CD-ROM release has been on the list matching risk. The level of detail of microdata now released via RADL and ABSDLs is substantially greater than is offered by the CD-ROM.

ABS microdata confidentiality practices have focused on these two aspects of spontaneous recognition and matching to lists for some time. For a number of reasons, the processes to address these risks has been subjective, long and tedious. This paper describes current and future work aimed to make these processes more objective and as automatic as possible.

The first Section of the paper deals with the issue of spontaneous recognition – how this is defined and how to protect data against it. The second Section discusses population modelling – the proposed technique for assessing both spontaneous recognition and the matching to lists risk. The third Section describes the applications of a population model in dealing with these confidentiality issues.

Since in the ABS the practice of making unidentifiable microdata accessible is much more prevalent and advanced for household surveys than it is for business surveys, the paper will primarily be written as if applying to household surveys (e.g. using terms such as individuals, persons and households). Many of the concepts and methods may also be applicable to surveys of businesses.

1. METHODS FOR PROTECTION AGAINST SPONTANEOUS RECOGNITION (SR)

1.1 Defining spontaneous recognition

Spontaneous recognition occurs when a user, while looking at information on a microdata file, recognises a particular record as possibly corresponding to a particular person that they know of. Looking at more information on the file for the individual confirms this identification.

One of the key assumptions behind the concept of spontaneous recognition is that it is an involuntary action. A user, in the course of their statistical use of the microdata file, comes across a record that seems to match the characteristics of someone they know. Because the action is involuntary, we consider that spontaneous recognition is only likely to involve a small number of data items – perhaps two or three – in combination at any one time. Simultaneously looking at more in combination would have to be a conscious act.

As an aside, much of the microdata confidentiality literature considers the scenario of an intruder or an attacker deliberately attempting to identify an individual from a microfile and thereby uncover information about that individual. We note that a conscious search for unusual combinations in the hope of identifying an individual would

i. contravene the undertaking the researcher signs in gaining access to the data; and

ii. be unlikely to succeed, assuming that the pool of 'very unusual' individuals the intruder knows is small (they are not matching to a list) and that few of these would have been sampled in the survey in question (the ABS avoids releasing unit data files that cover a high proportion of individuals of any given type).

While any user wanting access to ABS unidentifiable microdata must sign an undertaking which includes a condition that no deliberate attempt will be made to identify individuals on the file, spontaneous recognition could occur in the normal course of a researcher's activities. To quote Willenborg and de Waal (2001, p. 62):

> "It may happen that a researcher finds a very rare, or possibly even unique, combination of key variables by chance. This can happen, for instance, when the researcher examines a low-dimensional table. The curiosity of even the most innocent and well-respected researcher may be triggered when he comes across a very rare combination of key values in the data ... Because the researcher initially does not have the intention to disclose private information this disclosure scenario is also called spontaneous recognition."

1.2 A model for spontaneous recognition of a unit

Suppose that the intention is to deal with spontaneous recognition for each individual on a file. Consider the following events:

A. an individual included on the file is known by the user of the file;

B. the user happens to look at a particular set of data items from the record corresponding to the known individual;

C. the user notices that the values from this set of data items matches the characteristics of the known individual;

D. the user checks other data items on the file to verify a match; and

E. the individual is unique in the population based on all data items on the file that the user knows.

A model for the probability of spontaneous recognition of an individual by a given user could be as follows.

$$\begin{aligned}&\Pr(\text{spontaneous recognition for an individual})\\&= \Pr(A) \times \sum_{\text{all sets of data items}} \Pr(B \mid A) \times \Pr(C \mid B, A) \times \Pr(D \mid C, B, A) \times \Pr(E \mid D, C, B, A)\end{aligned}$$

The conditional probabilities recognise the fact that there is a sequence of events that need to occur.

This is a complex set of conditions that would be hard to evaluate individually; it would be very difficult if not impossible to obtain sensible estimates for all of these probabilities. Furthermore, some of the probabilities depend on the user, whereas we would like to confidentialise the file once to cover all users.

Instead of seeking an absolute measure of risk, we propose a measure of spontaneous recognition risk which depends on the proportions of units in the population displaying the particular combinations of data items seen in a given individual in the sample. For example, suppose an individual has a combination of age, income and occupation that appears in very few units in the population. This suggests that the person could be spontaneously recognised by a user who looks at these items for this individual, happens to know the individual, and notices the match. A subjective approach will be taken to determine the various combinations of data items that are likely to be recognised (i.e. looked at, known and noticed by a user), and to determine how few units in the population must display a given combination before the spontaneous recognition risk is unacceptable.

The approach could be expressed as follows. Consider an individual i in the sample, and a group of data items G. The combination G of data items constitutes a spontaneous recognition risk for unit i if fewer than c units in the population have the

same values for this combination as unit i does. The cutoff c would depend on which combination of data items is being considered – i.e. c is a function of G. This is discussed in more detail in the following two sections, although it is worth noting that if G consists of a large number of data items, we are assuming that $c=0$ – i.e. they are never a spontaneous recognition risk.

Note that we are still concerned ultimately about uniqueness in the population, although it is the recognisability of an unusual combination of a subset of these data items that is important for spontaneous recognition. Any unusual combination that is at risk of spontaneous recognition is likely to be unique in the population when only a small number of additional data items is considered. Of course, if there are a large enough number of data items on the survey file, there may be a considerable number of population uniques on it. To avoid such a situation would severely limit the detail of microdata that could be released. Rather, we limit access to users who have signed an undertaking not to attempt to identify any individuals and consider only small combinations of data items together, looking for individuals with values that are unlikely and recognisable.

In terms of which data items to consider, we have divided them into three sets. The first set is those data items which pose little or no identifiability risk since we consider that it is very unlikely that a user will know these characteristics about an acquaintance, or they are so mundane or common that they are not distinguishing. Examples would be expenditure on certain types of goods like bread or milk as collected in the Survey of Household Expenditure (expenditure on some larger items, such as cars, might be known), or distance travelled in their car in the last year. The second and third sets we have called *common knowledge variables* and *affinity variables*. Essentially, common knowledge variables are details which will generally be known fairly imprecisely for a relatively large group of people, while values of affinity variables will be known to a high level of precision for a subset of the population. They will be discussed in the following two sections.

1.3 Common knowledge variables

The concept of a *common knowledge variable* is that it is something about an individual that will be generally known by others, but not to a very fine level of detail. Examples are age, income and educational qualifications. For acquaintances, the age of a person would generally be known, perhaps not to the exact year but probably within a five year range. Again, general income level would be known to some extent. Similarly, educational qualifications of acquaintances could be expected to be known, perhaps at least whether they completed school, attended a tertiary educational institution or completed a higher degree. Even though these data items may contain greater detail than that described above, we propose that for spontaneous recognition purposes we would only need to consider them within these broader categories. This

is different to the practice of some overseas agencies e.g. Statistics Netherlands (Willenborg and de Waal 1996, pp. 51–62), but it important in increasing the detail of microdata that can be released.

Discussion point: Does MAC agree that a high level of detail of common knowledge variables does not need to be considered for spontaneous recognition?

Note that we are simplifying the knowledge that the user might have about individuals on the file. In reality, any person has most detailed knowledge of themselves, knows a lot of detail about close family and friends, has less knowledge about acquaintances (to varying degrees) and no knowledge at all about most of the population. The ABS does not legally need to protect against self-identification. Although the more detailed knowledge of close family and friends poses a higher identification risk, the number of acquaintances is generally far greater and so it is this level of detail of knowledge that we are assuming. Further, as indicated by the examples in the previous paragraph, it is not usually the fine detail of characteristics that makes a person unusual and recognisable, but unlikely combinations of broader data items.

According to the model proposed above, the objective is to determine what is an unusual combination of values from low order combinations (one-, two- or perhaps up to three-way) of these common knowledge variables. The main approach that has been used in the past is to calculate the number of people falling in each cell in one- or two-way tables of common knowledge variables from the sample. These cells are listed in ascending order of frequency, so that the rarest combinations appear at the top of the list. We are then concerned about any cells with frequencies lower than a certain threshold. This threshold corresponds to a general subjective assessment about how few people with certain common knowledge characteristics would there have to be to be considered unusual or rare. So far, we have been using a figure that corresponds to estimates of around 5000 or less, with more focus on the smaller estimates. Samples will generally supports estimates of this magnitude; the sampling weights for ABS household surveys are typically in the hundreds, so a population estimate of 5000 will correspond to around about ten contributors in sample.

Although we haven't formally done it yet, we think it would be worthwhile taking a sorted list of combinations and their frequencies as described (perhaps from the census rather than a sample), and getting feedback from a number of staff in the office about whether they considered the combination to be unusual. In this way, we could 'calibrate' such a threshold. This would replace the subjective assigning of a probability of the likelihood of recognition with an objective cutoff. The choice of

threshold would also be a balance between disclosure risk and damage to the data. A low cutoff would mean that almost all cells below it are unusual, whereas a high cutoff would mean that virtually no cells about it are unusual. In between is a mixture between unusual and not unusual – and probably a mixture of opinion as well!

1.4 Affinity variables

In addition to common knowledge variables which are generally known, there are a number of other variables – *affinity variables* – which may be known by people for a significant number of acquaintances to very fine levels. This is typically true where the user of a file has the same value for a particular characteristic as an acquaintance. The classic example is geography: people tend to know more about other people who live in the same neighbourhood, town or region. Other examples are country of birth, occupation, religion and sport played (or hobbies). People who are born overseas in a particular country are more likely than the general population to know other people in Australia who were born in the same country. Similarly, people of the same religion are more likely to know others of the same religion. Note that a variable could be both *common knowledge* and *affinity* (e.g. country of birth).

This detailed knowledge of an affinity variable, in combination with common knowledge variables, poses an increased identification risk. Because, for spontaneous recognition, we are only considering a small number of variables in combination at a time, we can restrict consideration of affinity variables to only one or two at a time in combination with common knowledge variables i.e. we are only considering three-way tables at most. In the original sequence of events leading to spontaneous recognition described above, the next step after noticing an unusual combination that corresponds to a known individual is to check other data items to confirm a match. This requires the user to make a judgment about whether such a check is worthwhile. If she/he considers that there may be 100 individuals in the population that have these characteristics, then she/he is unlikely to look further. On the other hand, she/he thinks there are only a handful (say four, for argument's sake) of individuals with these characteristics, then a 1 in 4 chance is probably high enough to be worth pursuing. Thus, because affinity variables are known to a fine level of detail, the threshold for any cells involving affinity variables is set at a much lower level – say, between 5 and 20 – than those consisting solely of common knowledge variables.

A sample will not support population estimates as low as 5–20. Therefore, to assess spontaneous recognition for combinations involving affinity variables, population data (e.g. census) needs to be used.

The above characterisation of spontaneous recognition differs with that used by some overseas agencies in several areas. Statistics Netherlands, for example, categorise data items into four class of identifiability: *extremely identifying*, *very identifying*,

identifying and *non-identifying*. Criteria such as *rareness*, *visibility* and *traceability* are used to assign data items to classes of identifiability. For example, *City of residence* is considered to be *extremely identifying*, as it is both very visible and very traceable (a user knows where to look for the individual); *sex* would be *very identifying*, as it is very visible, but not rare or traceable; *occupation* may be *identifying*, because only some values of occupation are rare or visible (Willenborg and de Waal 1996, pp. 51–52). Occupation is treated very differently in the two characterisations: in the Netherlands it is not considered a high risk, but under our characterisation it is an affinity variable and therefore could be likely to lead to identification. Conversely, our common knowledge variables will only be considered at broad levels, whereas in Statistics Netherlands they would be considered at the level of detail at which they are released.

Discussion point: Does MAC agree with the above characterisation of spontaneous recognition?

1.5 Treatment

Once a particular combination of common knowledge and affinity variables is determined as being a spontaneous recognition risk, the standard treatments are to mask the unusual records in some way by changing a value of one of the data items, or to collapse the categories of one of the data items. Masking is a local treatment, as only values of data items of particular records are changed, whereas collapsing is a global treatment as a data item's detail is decreased on all records on the file.

When should we use masking and when should we collapse categories? If there are a largish number of individuals in a particular category which pose an identification risk, it is generally better to collapse than to mask. There are two reasons for this: it is easier, and it is more transparent to users. It is more transparent, because the damage done to the data is explicit; it is the amount of detail that is lost by collapsing. In contrast, it is very difficult if not impossible for a user to assess the damage done by masking.

Affinity variables are often more effectively dealt with by collapsing, because they are by definition known to a high level of detail by those in the same class. Therefore, collapsing an affinity variable reduces the number of small '*affinity* by unusual' cells. If masking is to be used, it is often the value of the affinity variable that is useful to change, rather than that of another variable, in order to preserve the confidentiality of the individual or to create a false match rather than a true one.

For common knowledge variables, on the other hand, collapsing is not generally effective, as added detail does not increase the spontaneous recognition risk. An exception is where too much masking would be required; a common example is for extreme values, where it may be desirable to top code e.g. even if age is in single years, it is generally better to have a top category such as 80 years and older or 85+ rather than mask old people by reducing their age.

1.6 Complexities and practicalities

Household structure and data items

Effectively dealing with household structure and data items in terms of microdata confidentiality adds an extra layer of complexity to the problem. In considering spontaneous recognition for a hierarchical file, such as one containing households and multiple persons within households, it is necessary to consider unlikely combinations of individuals within a household, as well as unlikely data item combinations for a single individual.

The general approach is consider households as the units in an assessment for spontaneous recognition, with the data items of the individual members forming data items of the household. Even within households of the same size, it is clear that not all combinations of variables are likely to be viewed together. For example, the income of one household member and the country of birth of another may together be quite unlikely, but they may be unlikely to be viewed together. More likely to be viewed would be the same few data items for some or all of the members of a household, or a number of data items for an individual along with some 'household type' item (e.g. dwelling type, household composition) derived from the household information.

Similarly to the principle of only considering two or three variables in combination at a time for spontaneous recognition for persons, it is proposed to generally only consider two to three people within a household together. For example, the characteristics of a couple (a large age discrepancy may be unusual) or the age difference between mother and child (very small or very large differences may be unusual). The size of a household or family (i.e. a large number of children) could be unusual in itself, and so needs to be considered as a household characteristic.

Computing

The assessment above requires the number of units in the population taking various combinations of values. For common knowledge variables, this can be estimated based on the weighted survey data file. Aggregated weights are computed for all possible tables of up to three dimensions for a prescribed set of data items (the common knowledge variables). Adding an extra dimension by crossing with an affinity

variable requires the use of population data. The computer program to do these two steps needs to be very fast and efficient. The output would need to be ordered and presented in a useful way, for ease of clerical checking.

How to evaluate population counts for small cells

Since spontaneous recognition depends on the probability of the combination of unusual data items in the population, particularly when affinity variables are considered, some dataset containing information on the population is required. The weighted survey data file is one source; however, it is not useful for estimating extremely small population quantities.

The obvious and almost only alternative is the ABS Census of Population and Housing. The ABS has in fact used this file as the basis for evaluations up until now. While comprehensive in terms of coverage of people and households, the census is not ideal for this purpose for two main reasons.

The first is that the data items are typically not as detailed as those collected in the surveys which we are assessing. Some of the basic demographics are no problem, but for some data items in the census there is either a reduced level of detail (e.g. income only collected in ranges) or the data items may not be covered at all. In these cases, an approximation or proxy must be used.

Secondly, there is a timing discrepancy between the census and the surveys. The census may be up to five years out of date for a specific application. Luckily, for most data items, distributions are relatively slow to change and so what is an unusual combination is unlikely to vary much from census to census.

These difficulties lead to our current intention to move to using the survey data itself for the evaluations. The sample itself may be large enough to provide design-based estimates of combinations of common knowledge variables and thus fairly reliably identify unusual combinations – remember, we are expecting a threshold of around 5000. Unfortunately it will not support estimates involving a cross-tabulation by an affinity variable where we are talking about a threshold of 5–20. Population modelling based on the survey data is an alternative we have started to consider over the past several months, and the second part of this paper describes our thinking to date on this issue.

2. MODELLING POPULATION PROBABILITIES BASED ON SAMPLE SURVEY DATA

2.1 Assessing the risk of list matching

Section 1 of this paper has discussed spontaneous recognition, which is one way in which disclosure can occur. The second major source of possible identification from confidentialised unit record files (CURFs) released on CD-ROM is through matching to administrative lists. Protections built in to the remote access data labs and data labs at the ABS prevent users from matching. Thus the level of detail of microdata released via CD-ROM is substantially less than is offered by the other modes.

Released ABS CURFs are de-identified i.e. do not contain name, address or any other public-knowledge identifiers (a sequential identifier is placed on the files for easy reference). However, administrative lists often contain the same or similar demographic variables (age, sex, geographic location, etc.) as the ABS microdata. Datasets from the two sources could be linked via these data items and, in cases where the matches are one-to-one, the combined set of data items can pose an additional disclosure risk.

The ABS's Microdata Review Panel assessment of basic CURFs which are released on CD-ROM takes into account this identification risk from list matching. The general approach is to detect survey units that may be *population uniques*, whose identity would be able to be confirmed using the matched information from an administrative list.

Until the present, this risk has been simulated by matching the survey data to census files to determine population uniques in sample. This approach suffers from the same drawbacks as the use of census information outlined in the previous section. In fact, the age of the census data can lead to incorrect conclusions being drawn. While the same sorts of combinations that are unique tend to stay the same, the particular instances change over time e.g. by people growing older, changing occupations, etc.. It is therefore proposed to use the survey data itself to model the probability of particular combinations of data items appearing in the population.

Note that a similar method of population modelling is proposed to address disclosure risk via both spontaneous recognition and list matching. Like spontaneous recognition, for list matching we are only considering a relatively small number of data items, as the number of CURF variables likely to be on external lists is relatively small (although this is increasing due to improved data capture technology). However, the application in each context is slightly different; for list matching, we are interested in population uniques, whereas for spontaneous recognition we are interested in rare or unusual combinations. Further, for list matching the data items we are concerned

about are those on lists, whereas for spontaneous recognition we consider common knowledge and affinity variables - although there is a high degree of overlap in the two sets.

2.2 The confidentialising approach for list matching

The general approach to protect against list matching that is proposed is as follows.

i. Obtain a measure of the expected number of population uniques in the sample. Use this measure to assess which categories should be collapsed. Categories should be collapsed if they contain a large expected number of population uniques in sample.

ii. Identify (as best as we can) remaining records in sample which are likely to be unique in the population. What is likely to be unique is defined by a cutoff, which will be small but somewhat greater than one to allow for modelling error. Units in sample likely to be unique in the population will have some values masked so as to make them non-unique or false uniques.

2.3 Assessing population uniques given a population density function

Here is a proposed model for the population density. Let

$$\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{iK}\right)'$$

be a vector of K linking variables for unit $i = 1, \ldots, N$ in the population. Assume that the population of N units is generated from N independent and identically distributed draws from a superpopulation with a distribution defined by the density function $p(\mathbf{x}_i)$.

The number of times a combination of values of data items $\mathbf{x}_i$ appears in a realisation of the population will be $n(\mathbf{x}_i)$. This number has a Poisson distribution with mean $Np(\mathbf{x}_i)$.

Under this model, the probability that a particular combination $\mathbf{x}_i$ appears k times in the population is

$$\Pr\left(n(\mathbf{x}_i) = k\right) = \frac{e^{-Np(\mathbf{x}_i)} Np(\mathbf{x}_i)^k}{k!}$$

We are interested in the particular case where $k=1$, i.e.

$$\Pr\left(n(\mathbf{x}_i) = 1\right) = e^{-Np(\mathbf{x}_i)} Np(\mathbf{x}_i)$$

Turning now to the sample, not all population uniques will be in the sample – although all population uniques must also be sample uniques. The expected number

of population uniques in the sample, where the expectation is over both the model space and sample space, is

$$\sum_{i \in s} \Pr\left(n(\mathbf{x}_i) = 1\right) = \sum_{i \in s} e^{-Np(\mathbf{x}_i)} Np(\mathbf{x}_i)$$

where the sums are over units in the sample s. This is the first measure that we use. It is used to assess which categories should be collapsed. Once the collapsing is done, the process is repeated until the number of expected *population uniques* in sample is small for all combinations being considered.

For the second part of the process we use a test, of size α, of a combination $\mathbf{x}_i$ observed in the sample being a unique in the population, which is

$$Np(\mathbf{x}_i) < c_\alpha$$

The model of course would not be perfect, which is why a cutoff c_α is required, which would probably be some small number greater than one. It would be set to achieve the desired balance between type I errors (unique in population but expected number in sample greater than the cutoff c_α, i.e. $n(\mathbf{x}_i)=1$ and $Np(\mathbf{x}_i)>c_\alpha$) and type II errors (not unique in population but expected number in sample less than the cutoff, i.e. $n(\mathbf{x}_i)>1$ and $Np(\mathbf{x}_i)<c_\alpha$ over the superpopulation model). Using the Poisson density function, these type I and type II errors can be calculated as

$$\sum_{p(\mathbf{x}_i)>c_\alpha} \frac{1}{\pi(\mathbf{x}_i)} e^{-Np(\mathbf{x}_i)} Np(\mathbf{x}_i)$$

and

$$\sum_{p(\mathbf{x}_i)<c_\alpha} \frac{1}{\pi(\mathbf{x}_i)} e^{-Np(\mathbf{x}_i)} Np(\mathbf{x}_i)$$

respectively, where the $\pi(\mathbf{x}_i)$ are the sample selection probabilities. These statistics can be estimated from the sample by the sample sums

$$\sum_{i \in s, p(\mathbf{x}_i)>c_\alpha} Np(\mathbf{x}_i) e^{-Np(\mathbf{x}_i)}$$

and

$$\sum_{i \in s, p(\mathbf{x}_i)<c_\alpha} Np(\mathbf{x}_i) e^{-Np(\mathbf{x}_i)}$$

respectively.

In the above description, it has been implicitly assumed that the stratification variables are among the $\mathbf{x}_i$ s. This is not necessarily the case, but the derivation could be generalised by separating out a set of stratification variables $\mathbf{z}_i$.

Discussion point: The above method effectively uses the expected proportion of population uniques in the sample as the measure to assess disclosure risk. Some other international agencies use the probability of a unit being a population unique given that it is a sample unique (see, for example, Skinner and Holmes (1998)). Does MAC have any comments on the relative merits of using expected proportion of population uniques from the sample?

2.4 Modelling the population based on low order interactions

ABS samples, because of their size, will generally only be able to describe relatively low order interactions between variables. Population uniques will usually be defined by higher order interactions. To model the population from the sample, we will need some strong assumptions about how higher order interactions can be generated from low order interactions.

An implicit model for the population density $p(\mathbf{x}_i)$ can be specified based on a set of categories G that will correspond to low order interactions. We seek a set of probabilities for which

$$N \sum_{\mathbf{x}_i \in g} p(\mathbf{x}_i) = \hat{N}_g$$

for all categories $g \in G$, where $\widehat{N}_g$ is the estimate from the sample of total population falling into category g. Suppose that the total number of values that $\mathbf{x}_i$ can take in the population is small enough for us to enumerate the full set of corresponding probabilities $p(\mathbf{x}_i)$. Then the probabilities could be obtained by iterative proportional fitting (IPF), after setting some initial probabilities (say, assign equal probability to every possible combination of every value of the linking variables) and adjusting them iteratively so that they add to the appropriate totals for all the categories $g \in G$.

The IPF approach has the disadvantages that it is iterative, and it requires maintaining a large multidimensional array giving the probabilities corresponding to every combination $\mathbf{x}_i$.

2.5 An explicit parameterisation

An alternative to IPF is to take a parametric approach to modelling the population density $p(\mathbf{x}_i)$. An example is to consider the elements of $\mathbf{x}_i$ – the data items themselves – to each be a dimension of a multidimensional table which contains every K-way cross-classification from these data items. The cells could then be modelled explicitly – by a log-linear model, for example. In practice, we would look to simplify the model substantially by reducing the number of parameters, considering each data item individually as main effects, and perhaps some low order interactions that look important. Zero cells, including structural zeros, would cause a problem for a log-linear model, so some special treatment (e.g. a mixed model including a point mass at zero) may be necessary.

Whether to have an explicit or an implicit parameterisation is a key choice that needs to be made.

2.6 Modelling for larger sets of data items

Whether an implicit or explicit model is used, we will have difficulties if the number of values that $\mathbf{x}_i$ can take in the population becomes too enormous. Wedelin (1996) suggests that a modified IPF approach can still be done (approximately) in this situation; but it is not clear how the models can be used e.g. how can we determine the probability at some aggregated level under the model without enumerating the probability values for all the individual cells?

In cases with larger numbers of linking variables, the population modelling may need to be done separately for various subsets of the variables, rather than as a single large model for all the variables. This carries the danger that evaluations may not be entirely consistent with each other.

An alternative may be not to generate probabilities for all $\mathbf{x}_i$ values but for an appropriate sample of a very large number of values spread across the range of possible $\mathbf{x}_i$ values. For example, it may be possible to produce a 'generated population' that follows the modelled distribution (perhaps using Monte Carlo Markov Chain techniques). Then we could classify sample units as potentially population uniques if their values appear too few times in the generated population. The size of the generated population and the cutoff used could be adjusted to make the assessment appropriately robust at an acceptable computational cost.

2.7 Some notes on the form of density function envisaged

Regarding the form of the density function $p(\mathbf{x}_i)$, we envisage it being the product of low order interactions between data items. For example, suppose we are considering five data items, so that (dropping the subscripts i for the moment)

$$\mathbf{x} = \left(x_1, x_2, x_3, x_4, x_5\right)'$$

Then we might have a density function that looks something like

$$p(\mathbf{x}) = q_1(x_1, x_2)\, q_2(x_2, x_4, x_5)\, q_3(x_3, x_4)$$

i.e. the product of interactions between small subsets of the data items. Models of this kind arise naturally when conditional independence and causal relationships are present, and are known under the names of graphical models, log-linear models and Markov random fields. There is some literature available on how to select and estimate them (see, for example, Wedelin (1996)) although, because we are using them in an unusual context, we would appreciate any comments MAC has. One idea is that we could use the usual independence tests for contingency tables to determine which interactions need to be captured by the model e.g. in a two-way table, compare the cell counts with the relevant product of the marginals. Only two-way combinations that are not described well by independence would be included in the model.

Discussion point: Does MAC have any comments on model selection and fitting for graphical models to be used to predict population uniques?

2.8 Validation of the model

Data from the Census of Population and Housing will be used to validate the model. There are two parts to the evaluation: how well does the model fit, and how good are the methods in estimating it.

To assess the fit of the model, we would fit a low order model – the order being dictated by what could sensibly be estimated from a sample – to the full census data. The model predictions about the distribution of population uniques (frequency and characteristics) would be compared with the actual census data, which would tell us whether the graphical type model was giving a good description of the high level interactions present in the census data.

To assess the estimation, the procedure would be to take repeated samples, of sizes typical for those from which CURFs are produced, from the census data. We would go through the population modelling process described in the above sections for each of the samples generated, and check our results against the census data. We would be able to make this assessment directly for data items on the census, and we would need to extrapolate the results for other variables.

Discussion point: Does MAC have any comments on our proposed validation approach?

3. APPLICATION OF POPULATION MODELLING TO CONFIDENTIALITY ISSUES

3.1 Applying population modelling to household and business surveys

Household surveys

For household survey CURFs, we have recently developed standards for the levels of detail for data items usually included on every CURF. Previously, the areas running the survey tended to make their own judgments on what level of detail to release. Their decisions were affected not only by confidentiality considerations, but by data quality concerns e.g. avoiding categories where there were a small number of contributors because of the high sampling errors, collapsing detail due to coding or reporting concerns. As a result, the levels of detail tended to vary from collection to collection, which not only made the confidentiality assessment process longer (as a new detailed assessment needed to be made each time) but lead to inconsistencies for users.

The data items common to surveys, covered by the standards, tend to be the ones that are likely to appear on administrative lists (e.g. basic demographics, geography, income, occupation etc.) and therefore be at risk of leading to identification. What this means for population modelling is that it is probably only necessary to do the complete modelling process from scratch once every few years (say, after data from each census first becomes available). Between censuses, we would leave the set of predictors more or less fixed. It could still be necessary to re-estimate the model more often, perhaps annually or even from the survey data for each CURF. This would allow for some drift in the population distribution.

Business surveys

For business surveys, or other non-households surveys such as the Survey of Motor Vehicle Usage, producing CURFs is still an emerging area for the ABS. It is likely that in the short term at least, we will need to do intensive modelling for each CURF. There will probably be a need to undertake the confidentiality assessment quite quickly, and so there will be a requirement for fast computing for population modelling.

3.2 Treatment of population uniques

As with spontaneous recognition, the two main treatments will be to collapse (global recoding) or to mask (local value perturbation).

Most data items are categorical (or can be made so), and often there is a hierarchy in the values of the data items, e.g. age could be presented in single years, five-year

groups or ten-year groups. The collapsing approach would be group the data item up to the next level in the hierarchy if necessary, e.g. if there is a problem with one year olds, they could be grouped into a 0–4 years category. This results in a coarsening, $\mathbf{y}$, of the original data item/s $\mathbf{x}$. There is a relationship between the density functions of the two variables, i.e.

$$p(\mathbf{y}) = \sum_{\mathbf{x} \in \mathbf{y}} p(\mathbf{x})$$

If there is a confidentiality problem with the original set of data items, any necessary collapsing can be done and the population modelling can be repeated on the new set $\mathbf{y}$.

There may be several alternatives as to what to collapse. Any alternatives should, when applied to all units, reduce the expected number of population uniques below some threshold. We may be able to make the choice more objective may developing a measure/s of information loss, i.e. have some statistic which measures the amount of detail lost in the data by collapsing. We would then choose the collapsing which minimises the information loss. Ideally, we would like a process whereby this assessment and collapsing could be done automatically.

Discussion point: Does MAC have any ideas for best measures of information loss?

For masking, a similar process would be followed. The first step again would be to determine which the possible values to which the offending data item could be changed for the units in question which, when applied, would make it non-unique. The second step would be to randomly choose a different $\mathbf{x}$ value within the coarser $\mathbf{y}$ classification according to $p(\mathbf{x}_i)$. For example, if an individual is unique as a two-year old but not as a 0–4 year old, then randomly reassign their age to one of 0, 1, 3 or 4.

3.3 Other applications of population modelling

Evaluation of spontaneous recognition

As mentioned earlier, population modelling could be used instead of census data to estimate the number of people in the population in any two-, three- or four-way tabular cell. In particular, when we are considering an affinity variable crossed with an unusual combination of common knowledge variables, the design-based estimates will be too unreliable for the size of the estimates we are considering, and so population modelling could be used. In this case, the threshold that we set would be higher than

for list matching, as we are concerned about rare occurrences in the population, not just unique values.

The model would need to be sensitive enough to avoid being unnecessarily conservative. For example, a person aged over 65 by a couple of other characteristics may be rare in the Northern Territory, but not in South Australia. In this case, the model should pick up the fact that we need to change some characteristics of old people in the NT but not in SA.

Data cubes

The concept of a data cube is that it is a very fine multidimensional tabulation, from which all other tabulations likely to be required from a particular survey can be derived. An advantage of having data available via data cubes rather than from the original unit record data is that, if the data cube can be satisfactorily confidentialised, then any tables produced from it will be OK and will not need to be checked.

We have not yet considered the confidentiality of data cubes, although it is on our forward work program for 2004. One possibility may be to use the population modelling approach. We would need to set the level of significance, α, higher than we do to assess matching risk, as this will be the only protection applied (no undertakings would be signed).

Generating synthetic files for RADL

A third possible alternative use of population modelling is to generate synthetic files for the remote access data laboratory (RADL), to enable users to test out their programs before running them against the real files.

Currently, a test dataset for each expanded CURF available on RADL is produced. However, the data is essentially completely randomly generated, the only restriction being that legal values for each data are used. This file is made available to users via CD-ROM, so they don't have to test out their programs in the RADL environment (i.e. using the web). It is useful for users to test the syntax of their programs, but it doesn't enable them to determine whether their results will make sense, as the data bear little relation to the true data.

It would be helpful for users if the synthetic data have some of the same properties as the real data. For example, if distributions up to, say, three-way interactions are accurate. To do this, we would need to model potentially hundreds of variables – all of those on the CURF. To simplify our task, it may be possible to partition the data items into groups which are generally considered together, and preserve up to three-way interactions within these partitions. Another way of cutting down on the parameters would be to develop the model using the same data items as on the CURF

but using broader categories, and make assumptions about what happens at finer levels (e.g. use a linear model to interpolate).

Currently, this work has low priority and would probably only be progressed if there were high user demand.

ACKNOWLEDGMENTS

The material for this paper has been sourced almost exclusively from earlier thinking and documents by Bill Gross and Phil Bell. Bill and Phil also provided very helpful comments on an initial draft of this paper.

REFERENCES

Skinner, C.J. and Holmes, D.J. (1998) "Estimating the Re-identification Risk Per Record in Microdata", *Journal of Official Statistics*, 14(4), pp. 361–372.

Wedelin, D. (1996) "Efficient estimation and model selection in large graphical models", *Statistics and Computing*, 6(4), pp. 313–323.

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

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

FOR MORE INFORMATION...

INTERNET **www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications is available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries..

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice..

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney 2001

FREE ACCESS TO PUBLICATIONS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

2000001524404
ISBN 0 642 48164 4

RRP $11.00

Produced by the Australian Bureau of Statistics