Proposals for 2001 SARs: An assessment of disclosure risk

Angela Dale and Mark Elliot

CCSR Occasional Paper 16

The Cathie Marsh Centre

C C S R

for Census and Survev Research

Proposals for 2001 SARs: an assessment of disclosure risk

Summary

Marsh et al (1991) made the case for a sample of anonymised records (SAR) from the 1991 Census of Population. The case was accepted by the Office for National Statistics (then OPCS) and a request was made by the Economic and Social Research Council to purchase the SARs. Two files were released for Great Britain - a 2 per cent sample of individuals and a 1 per cent sample of households. Subsequently similar samples were released for Northern Ireland. Since their release, the files have been heavily used for research purposes and there has been no known breach of confidentiality. There is considerable demand for similar files from the 2001 Census, with specific requests for a larger sample size and lower population threshold for the Individual SAR. This paper re-assesses Marsh et al's analysis of the risk of identification of an individual or household in a sample of microdata from the 1991 Census and also uses alternative ways of assessing risk with the 1991 SARs. The results of both the re-assessment and the new analyses are reassuring and allow us to take the 1991 SARs as a baseline against which to assess proposals for changes to the size and structure of samples from the 2001 Census.

Proposals for 2001 SARs: an assessment of disclosure risk

1. Introduction

This paper sets out proposals for Samples of Anonymised Records from the 2001 Census. The proposals focus on the overall size and structure of the files and are based on experiences using the 1991 SARs. However, before assessing the risk to confidentiality of these proposals, we go back to the original research conducted by the ESRC Working Party on the 1991 SARs, led by Cathie Marsh, which provided the evidence upon which the risk of releasing the 1991 files was assessed. Marsh et al (1991) used a wide range of data to assess the risk of identification of an individual or household in a sample of microdata from the 1991 Census. More than 10 years on there is now much more evidence available – and, crucially, we have the 1991 SARs themselves.

Since 1991 there have also been advances in disclosure risk assessment for microdata. These include work on the data intruder's perspective (Mokken et al (1992), Lambert (1993), Dale and Elliot (1999)); the relationship between sample and population uniqueness (Fienberg and Markov (1996), Chen and Keller-McNulty (1998), Samuels (1998); and the impact of geographical detail on risk (e.g. Greenburg and Voshell (1990), Elliot et al (1999)).

We have therefore re-assessed risk of identification in the 1991 SARs using the same framework as Marsh et al (1991) (section 3) and also used alternative ways of assessing risk with the 1991 SARs (section 4). The results of both the reassessment and the new analyses are not only reassuring but also allow us to take the 1991 SARs as a baseline against which to assess proposals for changes to the size and structure of samples from the 2001 Census. These proposals are outlined in section 1.

2. User requirements for SAR for 2001

Extensive consultation with the user (and non-user) community, as well as the results of our own work on confidentiality, have resulted in proposals for changes to the Individual SAR to increase the sampling fraction and lower the population threshold.

The aggregation of local authorities was identified as a major problem for users of the 1991 SARs and prevented many others from attempting to use them (Brown and Dale, 1998). This was a particular concern for Scotland where sparsely populated areas required more aggregation than in the rest of GB. The move to Unitary Authorities will mean that, for 2001 SARs, only 39 per cent meet a population threshold of 120K. We estimate that 67 per cent would meet a 90K threshold and 90 per cent a 60K threshold. It is therefore a primary requirement to reduce this threshold to 60K if compatible with retaining confidentiality.

Secondly, there have been requests for a larger sampling fraction, particularly amongst those wanting to use the data at the Local Authority level but also to enhance the analysis potential of small groups, for example ethnic minorities.

Whilst there are a number of other improvements to the specification of the Individual SAR they are not seen as being fundamental to the overall specification of the files. Therefore this paper concentrates on the evidence by which these two proposals can be judged. However, we begin by re-assessing the evidence on which the release of the 1991 SARs was made.

3. Re-assessing risk of identification with the 1991 SARs

Marsh et al (1991) argued that disclosure of individual or household information from a sample of microdata first required that the individual or household was correctly identified[1]. They discussed the likely sources of an attempt at identification and based their empirical work on the likelihood of records from an external database being correctly matched with records from a SAR. Here we review the various steps set out by Marsh et al and re-assess the probabilities of each in the light of greater availability of data. The equation used by Marsh et al required three conditions for identification to be feasible and a fourth to be able to infer an exact record match. A probability was attached to each condition. These were:

1. An individual must appear in the sample. The probability of being in the sample was set at 0.02 – reflecting the 2 percent sample size of the Individual SAR.
2. An individual must be unique in the population on the key variables being used. The probability of being unique in the population on eight key variables was estimated at 0.02 based on 1980 census data for the entire region of Tuscany.
3. If a match is to be established between an individual in both the population and the sample, key variables must be recorded identically on both datasets. The probability of this occurring is was estimated at 0.6 using evidence from a wide range of datasets.
4. It is necessary to verify that a match is, in fact, a correct match and does not belong to someone else in the population with the same set of characteristics. The probability of verifying population uniqueness was estimated to be zero but set at 0.001 for the purposes of calculating a risk

We discuss each element in turn:

***Condition 1*:** The 1991 Individual SAR was a 2 percent sample of the enumerated population of Britain and therefore this probability is fixed at .02.

Condition 2: Probability of population uniqueness on a given set of keys.

Using data from the Italian census, Marsh et al established the extent of population uniqueness using the following key variables: age (single years), relationship to HOH

[1] Identity disclosure has been defined by Bethlehem, Keller and Pannekoek (1990:38): 'Identification of an individual takes place when a one-to-one relationship between a record in a released statistical information and a specific individual can be established'.

(4), sex (2), marital status (5), occupation (8), position at work (15), and tenure (3 categories), giving a theoretical total of 1.4 million cells. With geographical areas of approximately 100K population they found 2.4 per cent of records to be unique in the population.

We were unable to replicate this analysis exactly but used population data from the 1991 Census, at a geographical level of 120K population, and the following keys: age (94), sex (2), marital status (5), economic activity (5), ethnic group (10), migration (4), tenure (6), giving a theoretical total of 1.1 million cells. We found that 4.8 per cent of records were unique in the population.

It is likely that the higher percentage of uniques in the British data is explained by two factors:

1. the correlation between occupation and position at work in Marsh et al's data which will reduce the number of uniques and
2. the effect of two variables, ethnic group and tenure, in the British data, both of which are highly skewed and likely to increase the number of uniques.

The key variables we have used may represent a higher confidentiality threshold than that used by Marsh et al, in that accurate knowledge of all the characteristics would be very hard to come by. Nonetheless we use the higher figure of 0.048 rather than 0.02 in the re-calculation of Marsh et al's equation.

Condition 3: The probability of recording key variables identically in both datasets was based on:

- the amount of error and coder variability
- differences in the coding schemes used
- the difference in timing between two datasets

This is an area often neglected in work on confidentiality and yet it presents a very real, practical problem to the would-be intruder. To address this topic we conducted three pieces of empirical work each using a different data source:

1. An assessment of the extent to which the process of editing and imputation used to construct the 1991 Census database added a measure of protection to the data.
2. An analysis of change in key variables over time. This used the British Household Panel Survey which began in 1991 with a sample of 5,000 households across Great Britain who are interviewed each year. It therefore provided an excellent way of establishing the extent of change in key variables year on year.
3. The General Household Survey for 1991, with a sample of 12,000 households, contains most of the questions asked on the 1991 Census and was used to conduct a matching experiment with the 2 percent Individual SAR.

The results from this work are discussed below. First, however, it is important to remember an additional source of error introduced by the respondent at the point of completing the census return. This was addressed by the Census Validation Survey

(CVS) which assessed the extent to which census schedules had been incorrectly completed by the form-filler (Heady, Smith and Avery, 1996). Although in practice it is impossible to adjudicate between the correctness of answers on the Census and the CVS, the extent of difference in answers on the two occasions gives an indication of the gross error in responses to the 1991 Census questionnaire. We focus on those variables likely to be used as keys.

For a number of key variables there was little error – age and higher educational qualifications had a 2-3 per cent error rate and number of cars and tenure about 5 per cent gross error rate. However, gross errors were higher for economic position - nearly 8 per cent for men and nearly 14 per cent for women One of the most important key variables is occupation and here errors were as high as 25 per cent at occupational unit level and about 15 per cent at the level of the Standard Occupational Classification Major Groups (9 categories). The other important key variable is ethnic group. Overall, there was only 0.8 per cent error between the responses recorded in the census and the CVS; however when the white group is excluded the gross error rises to 13 per cent. The errors identified here are unlikely to have been subject to the edit and imputation procedures used in the 1991 Census unless they resulted in responses which were either out of range or inconsistent with related responses. They are, therefore, *additional* to those discussed in the rest of this section.

3.1 The effect of edit and imputation procedures

Population data from seven LAs drawn from the 1991 Census were made available by ONS under contract and with very tight security conditions. A pre-edit and imputation version of the data files and also a post-edit and imputation version enabled us to establish the extent to which this process added a measure of protection to the data. The ONS edit and imputation procedure is designed to check for consistency in a specified set of cross-tabulations - for example it specifies the valid values on marital status and economic activity for children under 16 - and if out-of-range values are found it imputes new values (Mills and Teague, 1991). It also imputes valid values for data items for which the respondent supplied no information and which are 100 percent coded.

In order to make this comparison we set up analyses using 14 standard key variables[2]:

Age (single year)
Sex (2)
Marital status (5)
Country of birth (42)
Ethnic group (10)
Long term limiting illness (2)
Primary economic activity (11)
Number of cars (3)

[2] Standard key variables are an organic set of variables defined by Elliot and Dale (1999) in their analysis of attack scenarios . They are those variables most readily available to a potential data intruder and coded at a level judged to be most reliable in terms of matching with the target file..

Central heating (3)
WC (3)
Bath (3)
Tenure (6)
Number of rooms (5)
Housing type (7)

All these census variables were coded at 100 per cent by ONS. We have considered only persons recorded as residents in households. For each of the seven areas for which we had data we compared the raw data as recorded at data entry with data after the edit and imputation processes. We recorded the number of records with different values on at least one of the 14 variables.

Table 1 shows the frequencies and percentage change[3] rates for each of the seven Local Authority areas. Overall, the percentage of records with one of more changes between the two files is over 16 per cent. There is considerable variation between the files, with local authority A having a much higher percentage of records changed than local authority G.

There is also variation in the number of variables on any one record that have changed between the two states. Table 2 shows the *number* of variables that have changed for each record, across all seven files. Of the 16 per cent of records with a change, nearly 80 per cent have changed in just one variable. In all 1.63 per cent of variable values across the data set changed in the edit and imputation process.

Table 3 gives an indication of whether the change in value was through imputation of missing data or editing of out-of-range or invalid responses. On average, about two thirds of the value changes between the raw and post imputation states are accounted for by the imputation of missing variables with the remaining third being due to editing of inconsistent or out-of-range values.

However, the proportion of the two types of change varies widely between variables. Marital status has the highest percentage of missing values – mainly associated with children under 16 – whilst in 5 per cent of records economic activity is subject to change through editing. There are also differences in the balance between missing and edited values. In the case of sex, virtually all the changes are due to missing values being imputed, whereas at the other extreme 87 per cent of changes to economic activity are due to editing procedures. In many cases imputation of missing values is likely to provide little additional masking as the imputed value may be self-evident - as in the marital status of a child under 16. However, for variables such as economic activity and housing type, both of which are widely available as a key variable, most change is through editing[4] and thus a significant amount of extra protection is added.

[3] We include change from a missing value to any value via imputation.
[4] 5.7 per cent of records are edited for economic activity and 0.5 per cent for housing type.

The foregoing analysis considers change averaged across all records. This begs the question: Are some kinds of people more likely than others to have records that are imputed or edited? If we find that there are systematic differences then we can identify those people who have more 'natural' protection from identification. From table 4 we can see that there is very little difference by sex in the percentage of changes recorded; however, single people have a much higher level of change than others. This is likely to be mainly associated with the under 16s. Those born outside the UK record a higher mean level of change as do all the non-white ethnic groups – particularly the Bangladeshis. Those who are recorded as students have mean percentage changes of almost 6 per cent. Looking at the variables associated with housing it is clear that those who share a bath and WC and who live in rented accommodation and in converted flats are more likely to have values edited or imputed.

The characteristics of these people are very similar to those identified as difficult to enumerate in the CVS and in most other surveys – students and single people living in multi-occupied housing. The high levels amongst the Bangladeshi population may suggest language difficulties in completing the census schedule.

From a disclosure point of view, however, the groups which have the greatest intrinsic protection are students, the unemployed and the long-term sick and those living in flats, particularly with shared facilities. These categories have the highest levels of edited records.

3. 2 The effect of data ageing

An additional source of protection comes from the fact that there is an inevitable time delay between the collection and release of census data. The SARs were not released by ONS until August 1993, two and a half years after data collection. During this time the circumstances of households and individuals change, thereby making more difficult any attempt to match records from 1991 with another data file or with individuals in the community. However, some characteristics such as sex are usually time constant; others such as ethnicity may be recorded differently at different time points. Age changes in entirely predictable ways whilst variables such as marital status are constrained in the options available (for example, one cannot move from single to divorced or vice versa). However, for many other variables there are no such constraints on the way in which change may occur and it is therefore much less predictable (change in area of residence, number of cars, economic activity).

Analysis of the British Household Panel Study for the years from 1991 to 1994, using variables recoded to maximise similarity with the 1991 Census, has provided some insights into the extent of change over time on different variables. Members of the study are re-interviewed each year and changes in status can therefore be recorded on a year on year basis.

Overall, 14 per cent of households in the study experienced a change in composition between 1991 and 1992 and almost half of this change was accounted for by the addition or departure of children (Buck et al, 1994). Other changes relate to the separation of

partners and the formation of married or cohabiting unions. However, the extent of household change varies considerably with the type of household. For example, the BHPS found only 1.4 per cent of households containing one elderly person changed composition whereas 27 per cent of lone parent households changed (op.cit).

Of those in employment at both 1991 and 1992, 18.5 per cent reported a change of job. Of these, 62 per cent changed their (detailed) occupation and 43 per cent changed their occupational group. A change of occupation was most likely for the least qualified, 72 per cent of whom changed their occupation when they changed job. Occupational change was greatest amongst sales workers and those in the least skilled jobs. Numbers in the survey are too small to allow an analysis by individual occupation, but one might expect that professional occupations such as doctor, dentist, pharmacist, solicitor, teacher, would be less likely to change occupation than others. Overall, about 10 per cent of adults moved house between 1991 and 1992, although many of these moves are within the same geographical area - town, or local authority (op.cit.). This evidence highlights the extent to which individuals change characteristics over even a one-year period. However, much of this change will be correlated – for example those who change their occupation may also move house.

We have conducted more detailed analyses to establish rates of change by variable between 1991 and 1992, 1993 and 1994. The following standard key variables were used:

- Standard region REGION12
- Marital status MLSTAT
- economic status JBSTAT
- Socio-economic group JBSEG
- Building type of accommodation HSTYPE
- Tenure of accommodation TENURE

Overall, only 2 percent of BHPS respondents changed region between 1991 and 1992, and 4.4 percent changed region between 1991 and 1994. Change was greatest in Inner London (8.4 percent and 17.2 percent, respectively) and least in Wales (table 4).

Table 5 shows considerable variation by type of housing. Greatest change is found amongst individuals living in converted flatlets and those living in business premises, whilst table 6 highlights the fact that private rented tenure is a transitory category. Marital status (table 7) shows that those who were separated in 1991 are unlikely to remain in this status for long. Similarly the unemployed, full-time students and those on government training schemes show the greatest rate of change between the economic status categories (table 8).

Occupational information, table 9, measured using socio-economic group, shows some groups which are particularly likely to record a change: small employers, foreman of manual workers, unskilled manual workers as well as those in the armed forces. For

these categories more than half respondents had changed status within two years. At the one-digit coding of occupation, SOC Major, only those in professional occupations had over two-thirds remaining in the same group by 1994. For most other groups percentages range from 46 (sales) to 63 (craft and related) (table 10).

Table 11 addresses the extent to which some of these variables are correlated. Using marital status, labour force status and tenure we find that only 68 per cent of those with a response in both 1991 and 1993 stay in the same categories on these three variables. Almost 20 per cent change both marital status and tenure, reflecting the extent to which forming or dissolving a partnership affects change between the major tenure categories – owner-occupation, social rented housing and privately rented.

By the time a dataset is two or three years old it is clear that there will have been considerable change on a number of key variables. In some categories of tenure and housing type change is particularly likely (e.g. rented furnished accommodation). Similarly other categories such as 'unemployed' are inherently unstable. The extent of change in employment status and occupation make them much less valuable as key variables than might initially be expected. In addition the difficulty of achieving accurate and consistent coding of occupation adds a further element of uncertainty when attempting record linkage.

It should be stressed that data ageing does not protect against a premeditated attack by an intruder using information collected close to the time of the census. However, Elliot and Dale (1998) found little evidence that date stamped information was routinely kept by holders of large databases although increasing data storage capacity means that caution must be exercised when considering this factor. Nevertheless, the effect of data ageing is a protection against opportunistic attacks. Furthermore, as the next section shows, even when identification information is collected close to the census, data divergence is still a considerable factor and the effects of data ageing may be considered as an extra layer of protection.

3.3 An experiment to match records from the GHS 1991 with the SARs 1991

Examining the extent to which individual records are subject to change, either through editing, imputation or ageing is important but, as Elliot (1998) argues, out this only considers errors on the target dataset and these may interact with those on the identification set. One way to understand this is to conduct matching experiments between datasets, such as those carried out on German data by Muller et al (1992). As there had been no similar matching experiment with British data, we designed an experiment to fill this gap and to address the question: what is the probability of achieving a correct match on a given set of keys using two datasets chosen to maximise comparability and thus the ability to match.

The experiment used microdata from the General Household Survey (GHS) and the Samples of Anonymised Records from the 1991 Census. These data were collected by

the same organisation - the Office for National Statistics (ONS) - for the same time period - 1991 - and included questions on same topics. Because the SARs are a 2 per cent sample with near-complete population coverage we can expect that 2 per cent of respondents to the GHS will be included in the SARs. In the case of the GHS this means that about 480 records might, by chance, be expected to be in both datasets.

The method of collecting the data is different between the two data sources – the GHS being interviewer based and the Census using self-completion. There are also some differences in the question wording and in the coding schemes used. However, we located 18 variables which were on both datasets and which could be recoded to be compatible. These 18 variables were therefore used as a key for matching. In some cases - sex - this was straightforward. In others - socio-economic group - comparability was much harder to achieve. The amount of time and effort required to achieve this set of key variables on both datasets was considerable - several months work were required to check definitions, set up recodes and do the necessary checking. This, too, was the experience reported by Muller et al. who attempted to match data from a social science survey with the 1 per cent Microcensus of the German federal state of North Rhine-Westphalia using up to 35 common variables.

When records from the GHS were matched against the SARs using the 18 variable key about 45 per cent of records in the GHS (over 11,000 records) gave an exact match against one or more individual in the SARs. In many cases there were very large numbers of 'statistical twins' in the SARs for one GHS record. (For example, 690 records in the GHS each had over 100 matches in the SARs). Thus for an intruder attempting to make a one-to-one match between the two data sets the situation would be very confusing. This very high level of false matches was entirely unpredicted and would seem to be quite an effective deterrent to the would-be intruder.

In an attempt to reduce the matching task to manageable proportions we focussed on one month only - April 1991 - designed to maximise the ability to match correctly. For this month there were 219 records in the GHS which matched one, and only one, individual in the SARs and a further 112 records which matched only two individuals in the SARs. From the sampling schemes we would predict that, by chance, there would be about 43 (the exact estimate is 43.22) people who were in both datasets.

From these results we can conclude that, for a user of an outside database, attempting this sort of match with no opportunity for verification would prove fruitless. In the first place, the small degree of expected overlap would be a considerable deterrent to an intruder. However, if a match between the two files was attempted the large number of apparent matches would by highly confusing as an intruder would have no way of checking correct identification.

When the 219 apparent matches were checked within ONS, only 6 were found to be correct and when the extra 112 records were checked a further 2 were correct.[5] From this we cannot obtain a direct estimate of how many of the 43 expected overlaps could be correctly matched as some potentially correct matches may have been amongst the records with more than two statistical twins in the SARs. We can conclude, however, that with data designed to maximise the likelihood of a match (recoding to the same classifications, using the same month, collected by the same agency and with similar questions), only 6 of the 43 expected matches could be correctly identified from one-to-one matches. Thus the probability of a correct match given a one-to-one match is 6/219=0.027. In practice, the reality faced by an intruder would be one of total confusion with 219 apparent matches and no way of knowing which was correct.

However, this probability does not directly address the requirement set by Marsh et al, in their third proposition. A measure of the probability of the two files being coded identically on the 18 key variables for those individuals expected to be in both files has been derived by Elliot using a disclosure intrusion simulation[6]. This shows that 68 per cent of records differed on at least one variable in the two datasets - even after they had been coded to maximise their compatibility.

However, the probability of recording key variables identically on two datasets is also influenced by data ageing. We saw evidence in table 11 from the British Household Panel Study that, using only three variables, 45 per cent of respondents present in both sweeps had changed their characteristics on one of more variables between 1991 and 1993.

These two figures – 32 per cent probability of being coded the same on both datasets and 55 per cent probability of changing values – are both conservative. Nonetheless, if

[5] These results are comparable to those obtained by Muller et al who expected 27 persons to be in both files but also found a much higher level of apparent matches (using 13 key variables over 1,000 cases in the Microcensus file had one or more matches in the social science survey and with 21 variables this was 298) and achieved no correct identifications.

[6] The data intrusion simulation method (DIS) developed by Elliot (1988, forthcoming) enables the estimation of the probability of a correct match given a unique match between any hypothetical external file and a particular target file with zero data divergence (no difference in variable values for the same individual).

By comparing the matching probability generated by DIS with the actual probability found in the matching experiment, we can obtain a measure of the effect of data divergence between the SARs and GHS on the matching success rate obtained in the experiment. DIS was run on the SAR with same key variables and level of geographical detail as was used in the matching experiment. The estimated probability of a correct match using a unique match generated using this method was 0.088. This compares with 6 true matches from 219 unique ones actually found in the experiment, a probability of 0.027. This indicates a *single variable data divergence rate* between the two files of:

$$1-(0.027/0.0.88) = \sim 0.68$$

In other words approximately 68 per cent of records differed in how one or more variables had been recorded in the final files.

taken together they suggest that a probability of 0.18 (.32 * .55= 0.18) can be used as a more realistic estimate than that of 0.6 used by Marsh et al.

We have presented evidence of inconsistency in response, editing and imputation and change over time in key variables. In addition, the matching experiment demonstrated the difficulties, in reality, of achieving a match in circumstances that maximised the likelihood of matching.

It is important to recognise that the difficulty of matching records in the GHS and the SARs was largely because of the high degree of correlation between key variables. This is partly because almost all the key variables are categorical. Work in the USA (Winkler, 1999) has shown that the level of matching is much better when income data is used as part of a key to match against administrative tax records. This is largely because income is recorded as a continuous variable in the US census and the response given on the census schedule is likely to be taken from the respondent's self-completion tax return. An income question will not be added to the 2001 Census and therefore this additional, and potentially more successful, basis for matching will not be available.

Condition 4. Marsh et al set the probability of verification at 0.001, although they argued that, in reality, it was more likely to be zero.

Although, ten years on, we have seen a huge increase in the compilation of lifestyle databases, and in the power of computers to search these databases, these do not provide a basis for verifying that any given record is unique in the population. Lifestyle databases do not give population coverage; they do not record historical data allowing matches to a given year, let alone month; they do not record information using the same classifications as the census; they are renown for their inaccuracy. Even without these data quality limitations, lifestyle databases are designed to hold the most up to date information possible on each individual rather than keeping information relating to a specific period some years earlier. They could not, therefore, provide *verification* that an apparent match was a correct match.

We do not yet have a complete population register. Data linkage within government is still proceeding cautiously; there is no available register with the range of variables needed to verify identification. In addition, were such a register to be compiled it would be kept under heavy security and accessed only by a limited number of staff screened for security.

It may well be argued that verification would not be necessary in order to damage the census. However, in order to rework Marsh et al's figures we continue with the verification requirement and reach the conclusion that scope for verification remains very limited. We therefore use the same probability as Marsh et al, 0.001.

3.4 Calculating the per record risk:

From the four element discussed above, Marsh et al calculated a per record probability of identification of 2.4×10^{-7} based on probabilities of:

0.02 (sampling fraction) *
0.02 (uniques in the population) *
0.6 (probability of achieving a correct match) *
0.001 (probability of verifying a match)

If we re-calculate this probability using our more recent figures we reach a probability of : $.02 \times .048 \times .18 \times .001 = 1.73 \times 10^{-7}$

(.02 giving the probability of a record being in the sample; .048 the probability of a record being unique in the population; .18 the probability of achieving a correct match; and .001 from the probability of verifying such a match).

From this re-working of the evidence used to judge the case for the 1991 SARs we can conclude that the per record risk of identification calculated by Marsh et al (1991) under-estimated the difficulty of making a correct match between the SARs and an external data file. In particular, they did not have available longitudinal data to establish the extent of year on year change in key variables; neither were they able to conduct the kind of matching experiment with another dataset that highlighted the extent of false matches found with the 1991 SARs and the 1991 GHS.

3.5 The disclosure risk from changes in sample size and population threshold

From the revised probabilities we can calculate the increased risk of a larger sample size:

Increasing the sample size to 3 per cent gives:

$.03 \times .048 \times .18 \times .001 = 2.59 \times 10^{-7}$

Thus the effect of increasing the sample size from 2 per cent to 3 per cent gives a per record risk very similar to that accepted for the 1991 SARs.

We can also calculate the effect on the percentage of population uniques of reducing the population threshold. For the same key variables, the percentage of population uniques rises to .054 at a population size of 90K giving a per record risk of::

$.03 \times .054 \times .18 \times .001 = 2.92 \times 10^{-7}$

Reducing the population threshold to 60K gives:

$.03 \times .067 \times .18 \times .001 = 3.62 \times 10^{-7}$

From this we can argue that reducing the population threshold to 90K and increasing the sample size to 3 per cent leaves the per record risk very similar to that accepted for the 1991 SARs. Reducing the population to 60K increases the risk slightly by comparison with 1991.

It is, however, necessary to point out that increasing the sample size increases the total number of records that are at risk, even if the risk per record does not increase substantially.

3.6 Conclusions on the disclosure risk of the 1991 Individual SAR

In conclusion it is important to remember that:

The equations used above are based on the assumption that an attempt at identification is made. If we included in the probability equation the likelihood of an attempt, the probability would reduce even further.

There has been no known attempt at identification with the 1991 SARs – nor in any other countries that release samples of microdata. Our research on the scenarios under which an attempt might occur suggests that there would be no commercial advantage in attempting to make direct matches with an external database and that the main danger comes from maverick attempts to discredit either the census operation, ONS or the government (Elliot and Dale, 1999). Attempts to discredit these bodies would be much simpler using more readily accessible data – for example, by infiltrating the census-taking process. Obtaining a copy of the SARs and then going through the time-consuming process of attempting identification requires considerable expertise at data manipulation and as well as time.

It has been possible to evaluate the confidentiality work which underpinned the release of the 1991 SARs because of the recent availability of a much greater range of data – including the 1991 Census itself. In the following section we take advantage of this greater availability of data to use some alternative methods of assessing the risk of the 1991 SARs and proposals for 2001.

4. Alternative ways of assessing risk

Access to population data from the 1991 Census has allowed us to develop some alternative methods of risk assessment. These methods all draw samples from the population data and therefore there is no disparity between the population and the sample in terms of classification, editing, or data ageing. All results are based on individuals in households for one large Local Authority District.

We begin from the same premise as Marsh et al: that records at risk of identification are those which are unique in both the sample and the population. These records are termed union unique (UU) in the rest of this discussion. Similarly sample uniques are termed

SU and population uniques PU. We draw samples of different size and with different population thresholds in order to compare the absolute and relative numbers of records unique in both sample and population (UU). We can also make these comparisons using different key sets.

There are various ways in which we can express the number of records that are unique in both the sample and the population.

1. The percentage of Sample Uniques which are Union Unique (UUSU)

$$\frac{UnionUniques}{SampleUniques} \text{ x } 100$$

This gives us an estimate of the likelihood that those individuals unique in the SARs are also unique in the population. If we increase the UUSU then we may be concerned that a higher percentage of our sample uniques are in the risk set. This measure has been used to assess the effect of changing sample size or population threshold (Elliot, Skinner and Dale, 1998). It has also been used by Carter (1991) in work with Canadian data.

2. An alternative way of expressing the UU risk set is as a percentage of the entire sample. This gives a more direct measure of the percentage of a given sample that is at risk.

3. We can also calculate the absolute numbers in the UU risk set and establish how this changes with sampling fraction and population threshold.

For each measure we can use the value that would be obtained for a sample with the parameters of the 1991 SAR: that is, a 2 per cent sample at 120K population threshold. Having provided evidence that the 1991 SARs were rather safer than assumed when ONS agreed to their release, we can now go on to use them as the base-line against which to compare alternative sample specifications. Each measure gives us related, but different, information on our risk set. In order to clarify this we have calculated each measure using data at three geographical levels (60K, 90K and 120K). At 120K the LA was divided into four subdivisions; at 90K there were five and at 60K there were seven subdivisions. Each subdivision was built up from wards that were geographically contiguous and broadly homogenous in social and demographic characteristics.

For each geographical area we have used a 2 per cent and 3 per cent sample and two different key sets: a basic 3-variable key and a 5-variable key. Results represent mean values across all the areas within the LA District.

4.1 **The percentage of sample uniques which are union unique (UUSU)**

Table 12 shows UUSU values at different sampling fractions and different geographies for each key set. We can see that there is little increase in the UUSU as the size of the

population threshold decreases – particularly with the larger key (table 1b). An increase in sample size does, however, increase the UUSU value. Although the sampling fraction has increased by 50 per cent the increase in the UUSU value is around 30 per cent in table 1b. In table 1a, with the small key set, there is considerable variation in the increase in the UUSU value with the larger sample. A number of other interesting differences with the key sets used are explored in more detail elsewhere (Elliot et al, 1998).

Based upon this evidence we can see that, using the larger and more realistic key set, decreasing the geographical threshold to 60K has very little impact. This counter-intuitive effect is found using other key sets (Elliot et al, 1998) and may be explained by the clustering of individuals on a range of descriptive variables. Unusual individuals (or 'special uniques') with idiosyncratic combinations of variables (for example, widowed 16-year olds) are likely to be union unique at any level of geography. However, with larger keys, the correlated characteristics of individuals (evident in the matching experiment with the SARs and the GHS) ensures that, even at lower levels of geography, most individuals have a statistical twin in the population. This finding is explored further, and a stochastic model fitted to explain the observation, in Elliot, Skinner and Dale (1998).

4.2 **UU as a percentage of the sample size**

The number of union uniques can also be expressed as a percentage of the total sample and thus gives an alternative way of assessing the size of this risk set.

Union uniques, using these two sets of keys, form a very small percentage of the total sample size. For the larger key the relationship with the population size is not proportional – thus an increase in population threshold from 60K to 120K leads to only a 50 per cent increase in the percentage union uniques. By contrast, using the smaller key set the increase in union uniques is proportional at a 3 per cent sampling fraction and disproportionately high with a 2 per cent sample. Again, the reasons relate to the role of 'special uniques' in the small key set and the effect of correlation between individuals with the larger key set.

4.3 **Number of union uniques**

The third measure used here is the total number of records (per geographical unit) which are union unique. This gives an absolute measure of the number of records at risk within each geographical area defined.

We can see that as the key size increases the number of union uniques increases considerably. However, the relative change with sample size and population threshold is the more relevant consideration. As the population size decreases, the number of UUs decreases – a function of a reduction in the number of both sample and population uniques. However, the increased sampling fraction acts to increase the number of sample uniques and thus increases the number of UUs. With the large key set (table

14b) this is approximately proportional to the increase in the sampling fraction. Thus with a threshold of 120K and a 2 per cent sampling fraction there are 21 risky records per geographical unit which increases to 26 per geographical unit with a larger sampling fraction and smaller population threshold.

The comparisons between geographical level shown above have been calculated taking each area as a single entity and have not considered the effect of using different thresholds on the disclosure risk for the entire sample for GB. The aggregation effect can be simply demonstrated with the LA data used for these examples.

The total population was divided into seven areas of about 60K, five at 90K and four at 120K. Using the numbers from table 14b which average across the constituent areas, we can see that the total number of records which are union unique in the Local Authority District with a 2 per cent sample are:

60K 17.8 x 7 = 125
90K 21.0 x 5 = 105
120K 20.7 x 4 = 83

Once again, the increase in the number of individuals who are union unique increases by 50 per cent for a halving of the population size. In assessing the significance of this, a crucial question is the role of geography in a disclosure attempt. If it is assumed that geography is likely to be the most important key variable, then one might argue that it is more important to assess risk within a defined geographical area, either at the LA level or below, rather than to aggregate risky records across the entire country.

5. Conclusions

The 1991 SARs represented a major breakthrough for social science. They have been widely used and there is every indication that the 2001 SARs will be used even more widely. However, there are powerful arguments for increasing the sample size and reducing the population threshold of the Individual SAR. User consultation has demonstrated that these changes would make a significant impact on the research value of the SARs and widen the user base considerably.

In this paper we have re-assessed the per record disclosure risk of the 1991 SARs and shown that it is more likely to have over-estimated the risk than under-estimated it. We have gone on to show that, using a number of different measures, an increase in sample size to 3 percent and a reduction in population threshold to 60-90K makes only a small additional increase in the risk to confidentiality.

There have been no known attempts to breach confidentiality in the 1991 SARs and a rigorous registration system requires users to give a legally-binding undertaking not to attempt to identify any individual or household. Scenario-based analysis of threats to confidentiality (Elliot and Dale, 1999) suggest that these are more likely to occur during

the process of census-taking when public awareness is higher and opportunities are easier than from attempts to identify individuals in microdata files.

The value of the research and policy value of the 1991 SARs has been very considerable. The proposed increase in sampling fraction and reduction in population threshold would increase it even further and broaden the value of the SARs in the policy arena.

References

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990) 'Disclosure control of microdata' in *Journal of the American Statistical Association, 85*, 409, pp38-45

Brown, M. and Dale, A. (1998) A Survey of User Requirements, *SARs Newsletter* No.11, September 1998

Buck, N., Gershuny, J., Rose, D. and Scott, J. (1994) *Changing Households, The British Household Panel Study, 1990-1992*, Essex: Centre for Micro-Social Change

Carter, R., Boudreau, J.-R, Briggs, M. (1991): Analysis of the Risk of Disclosure for Census Microdata. *Statistics Canada Working Paper*.

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

Elliot, M.J. (1998) DIS: Data intrusion simulation - a method of estimating the worst case disclosure risk for a microdata file. *Proceedings of international symposium on linked employee-employer records*, Washington; May 1998.

Elliot, M. J., and Dale, A. (1998) Disclosure Risk for Microdata.: Workpackage DM1.1 What is a Key Variable? *Report to the European Union ESP/ 204 62/DG III*
.
Elliot, M. J. and Dale, A. (1999) Scenarios of Attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*. Spring 1999.

Elliot, M. J., Skinner, C. J, and Dale, A. (1998) Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*; 1(2)

Fienberg, S. E. and Makov, U. E.(1996), Confidentiality, Uniqueness and Disclosure Avoidance, for categorical Data. *Proceedings of 3rd International Seminar on Confidentiality*. Bled, Slovenia, 165-174.

Greenburg, B.; Voshell, L. (1990) The Geographic Component of Disclosure Risk for Microdata, American Bureau of the Census *SRD Research Report* Census/SRD/RR-90/13.

Heady, P., Smith, S. and Avery, V. (1996) *1991 Census Validation Survey: quality report*, London: HMSO

Lambert, D. (1993) Measures of Disclosure Risk and Harm. *Journal of Official Statistics, 9(2) 313-332*

Marsh, C.; Skinner, C.; Arber, S.; Penhale, P.; Openshaw, S.; Hobcraft, J.; Lievesley, D.; Walford, N. (1991). The Case for a Sample of Anonymised Records from the 1991 Census. *Journal of the Royal Statistical Society* Series A,154, 305-340.

Mills, I. and Teague, A. (1991) 'Editing and imputing in the 1991 Census', *Population Trends, 64, 30-37*

Mokken, R. J., Kooiman, P., Pannekoek, J. and Willenbourg, L. C. R. J. (1992) Disclsoure Risks for Microdata. *Statisica Neerlandica*, 46, 49-67.

Muller, W; Blien, U., Wirth, H. (1992) Disclosure risks of anonymous individual data. *Paper presented at International Seminar for Statistical Disclosure*, Dublin, September 1992.

Samuels, S. M.(1998). A Bayesian Species-Sampling-Inspired Approach to the Uniques problem in Microdata, *Jounal of Official Statistics,* 14(4), 373-384.

Winkler, W. (1999) Re-identification methods for evaluating the confidentiality of analytically valid microdata, *Research in Official Statistics, 1(2), 87-104.*

Table1: Percentage of records of persons in households on each LA file where one or more of the 14 variables changed during the edit and imputation process

File	*Number of records*	*% changed*	*% no match*
A	216740	24.10	0.22
B	445355	17.18	0.06
C	190831	12.99	0.18
D	66040	11.85	0.75
E	31930	14.26	0.12
F	129262	14.35	0.09
G	76316	7.29	0.53
All	1156474	16.43	0.19

Table 2: Number of changes per record between raw and post-imputation files.

number of changed values	*Frequency*	*percentage of all records*	*percentage of changed records*
0	966454	83.57	-
1	150241	12.99	79.07
2	25581	2.21	13.46
3	6613	0.57	3.48
4	2841	0.25	1.50
5	1783	0.15	0.94
6	1142	0.10	0.60
7	725	0.06	0.38
8	385	0.03	0.20
9	281	0.02	0.15
10	198	0.02	0.10
11	74	0.01	0.04
12	122	0.01	0.06
13	32	0.00	0.02
14	2	0.00	0.00

Table 3: Frequency and percentage of records with imputed and/ or edited values for each of the key variables

Variable	i. No. missing values imputed	% of records	ii. No. edited values	% of records	i. and ii. combined	%
Sex	4436	0.38	157	0.01	4593	0.40
Age (94)	9001	0.78	3260	0.28	12261	1.06
Marital status (5)	41101	3.55	658	0.06	41759	3.61
Country of birth (42)	8569	0.74	2414	0.21	10983	0.95
Ethnic group (10)	14988	1.30	39	0.00	15027	1.30
LTL illness (2)	14658	1.27	3751	0.32	18409	1.59
Economic act. (11)	10084	0.87	65547	5.67	75631	6.54
Cars (4)	15042	1.30	266	0.02	15308	1.32
Central heating (3)	10554	0.91	153	0.01	10707	0.93
WC (3)	6413	0.55	2521	0.22	8934	0.77
Bath (3)	5877	0.51	2386	0.21	8263	0.71
Tenure (6)	10638	0.92	299	0.03	10937	0.95
Rooms (5)	22241	1.92	1339	0.12	23580	2.04
Housing type (7)	2199	0.19	5168	0.45	7367	0.64

Table 4: Mean % changes between raw and post imputation states for each category of the key variable

Variable	Value	Mean % changes
Sex	Male	1.60
	Female	1.56
Marital Status	Single	1.78
	Married	0.77
	Re-married	0.66
	Divorced	0.90
	Widowed	1.26
Country of Birth	UK	1.36
	Other	2.25
Ethnic group	White	1.11
	Black Caribbean	3.46
	Black African	4.29
	Black Other	3.35
	Indian	2.55
	Pakistani	3.81
	Bangladeshi	5.07
	Chinese	2.78
	Other Asian	2.84
	Other other	2.81
Long term limiting illness	Yes	1.13
	No	1.37
Primary Economic Status	Child	1.90
	Employee full-time	0.68
	Employee part-time	0.91
	Self Emp. With employees	1.11
	Self Emp without employees	0.68
	Govt. Scheme	1.34
	Unemployed	1.42
	Student	5.87
	Permanently Sick	1.73
	Retired	0.88
	Other Inactive	0.98
Number of Cars	None	1.87
	One	1.28
	Two	0.99
	Three or more	0.84
Central heating	All rooms	1.33
	Some rooms	1.17
	No rooms	1.43
Inside WC	Exclusive	1.42
	Shared	8.58
	None	1.53
Bath	Exclusive	1.43
	Shared	6.03
	None	1.88
Tenure	Owned Outright	1.29
	Owned buying	0.98
	Rented with job or buiness	1.49
	Rented from LA/HA etc	1.93
	Rented privately furnished	1.91
	Rented privately unfunished	1.71
Number of Rooms	1-3	1.71
	4-6	1.38
	7-9	1.38
	10-12	1.44

	13+	1.63
Housing	Non-permanent	1.81
	Unshared detatched	0.93
	Unshared semi	1.22
	Unshared Terraced	1.84
	Unshared flat residential	1.95
	Unshared flat commercial	2.12
	Converted flatlets and non self contained spaces	2.69

Table 5: Percentage of BHPS respondents with same response in 1992,1993 and 1994 as that given in 1991 for HSTYPE (housing type).

Value Label	***1992***	***1993***	***1994***
Other	0.0	0.0	0.0
Detached house	87.1	80.9	81.9
Semi-detached house	82.9	76.7	73.4
End Terrace house	57.9	51.6	49.8
Terraced house	75.0	68.0	67.7
Purpose built flat	83.5	70.7	65.4
Converted flatlets	66.4	52.0	41.5
Inc business premises	62.7	56.4	42.9
Bedsit multi occupancy	0.0	0.0	0.0
Bedsit other	33.3	0.0	0.0
Total	79.0	71.7	69.6

Table 6 Percentage of BHPS respondents with same response in 1992,1993 and 1994 as that given in 1991 for tenure

Value Label	***1992***	***1993***	***1994***
Owned Outright	93.7	88.3	89.1
Owned with mortgage	93.9	87.3	85.5
Loacl Authority	92.4	85.4	81.9
Housing association	74.0	74.6	72.3
Rented from employer	65.6	75.9	63.9
Rented privately unfurnished	66.3	51.3	45.6
Rented privately furnished	65.6	45.9	38.5
Rented other	0.0	0.0	0.0
Total	90.8	84.0	82.1

Table 7 Percentage of BHPS respondents with same response in 1992,1993 and 1994 as that given in 1991 for marital status

Value Label	***1992***	***1993***	***1994***
Married	97.5	95.5	93.8
Separated	58.0	37.0	30.3
Divorced	91.7	85.6	82.5
Widowed	97.7	98.4	97.4
Never Married	95.1	89.4	85.7
Total	95.9	92.6	90.3

Table 8 Percentage of BHPS respondents with same response in 1992,1993 and 1994 as that given in 1991 for economic status

Value Label	***1992***	***1993***	***1994***
Self Employed	78.5	70.7	66.9
In paid employment	87.6	83.2	81.3
Unemployed	46.3	36.7	27.9
Retired	87.7	90.5	90.6
Family care	72.4	61.7	58.8
FT student	68.0	51.5	40.8
Long term sick	70.4	75.0	68.9
On maternity leave	0.0	0.0	10.0
Govt. trng scheme	19.4	0.0	12.5
Other	10.0	3.4	3.4
Total	81.2	76.6	74.1

Table 9 Percentage of BHPS respondents with same response in 1992,1993 and 1994 as that given in 1991 for socio-economic group

Value Label	***1992***	***1993***	***1994***
Employers,Large	71.4	66.7	83.3
Managers Large	60.6	54.8	53.1
Employers Small	67.3	60.3	56.7
Managers Small	47.0	41.3	40.2
Professional self employed	73.5	66.2	61.9
Professional employed	63.4	53.3	54.1
Int non-manual	64.6	60.8	56.2
Int non-manual foreman	40.2	34.6	32.5
Junior non-manual	70.3	61.0	56.4
Personal Services	63.8	57.1	48.2
Foreman manual	48.6	46.0	37.8
Skilled manual	69.0	56.8	54.9
Semi-skilled manual	63.5	49.9	50.7
Unskilled manual	52.0	42.7	41.7
Own account workers	66.2	53.1	51.3
Farmers employers	55.6	70.6	56.3
Farmers own account	71.4	57.1	46.7
Agricultural workers	75.0	53.6	62.3
Armed forces	25.0	20.0	20.0
All	62.7	54.3	51.3

Table 10 Percentage of BHPS respondents with same response in 1992,1993 and 1994 as that given in 1991 for SOCMAJOR.

Value Label	***1992***	***1993***	***1994***
Managers & administrators	69.6	70.5	61.9
Professional Occupations	75.9	70.9	69.1
Associate professional & technical	68.8	65.0	61.7
Clerical & Secretarial	74.5	67.7	61.8
Craft & related	74.5	66.6	62.5
Personal & protective	72.1	65.5	59.3
Sales	62.9	50.4	45.7
Plant & Machine operatives	69.9	62.9	58.1
Other	60.2	52.1	50.4
All	70.7	64.7	59.9

Table 11 Extent of change for BHPS respondents between 1991 and 1993 on three variables: marital status, labour market status and tenure

Change		Frequency	Percent	Valid Percent
No change	.00	5232	54.5	67.6
Marital & lab mkt	1.00	322	3.4	4.2
Marital & tenure	2.00	1495	15.6	19.3
Lab mkt & tenure	3.00	349	3.6	4.5
Lab mkt only	4.00	62	.6	.8
Marital only	5.00	134	1.4	1.7
Tenure only	6.00	123	1.3	1.6
Marital, lab mkt & tenure	7.00	27	.3	.3
	Total	7744	80.7	100.0
Missing	System	1856	19.3	
Total		9600	100.0	

Table 12 UUSU by sampling fraction and geographical level

a. Key set: age (94), sex (2), marital status (5)

Level	60K	90K	120K
2%	0.56	0.43	0.37
3%	0.72	0.77	0.59

b. Key set: age (94), sex (2), marital status (5), economic activity (5), ethnic group (10)

Level	60K	90K	120K
2%	6.20	6.21	6.04
3%	8.04	8.22	7.84

Table 13 UU as a percentage of sample size by sampling fraction and geographical level

a) Key set: age (94), sex (2), marital status (5)

Level	60K	90K	120K
2%	0.06	0.04	0.02
3%	0.06	0.04	0.03

b) Key set: age (94), sex (2), marital status (5), economic activity (5), ethnic group (10)

Level	60K	90K	120K
2%	1.40	1.18	0.93
3%	1.38	1.17	0.91

Table 14 Number of Union Uniques by sampling fraction and geographical level

a) Key set: age (94), sex (2), marital status (5)

Level	60K	90K	120K
2%	0.8	0.6	0.6
3%	1.1	1.1	0.8

b) Key set: age (94), sex (2), marital status (5), economic activity (5), ethnic group (10)

Level	60K	90K	120K
2%	17.8	21.0	20.7
3%	26.4	31.2	30.6