

## **Request to ONS for Samples of Anonymised Records from the 2001 Census**

Angela Dale,  
CCSR,  
University of Manchester

26th September 2001

### **1. Introduction**

In 1990 the Census Offices agreed to release microdata samples from the 1991 Census. An ESRC Working Party, led by the late Cathie Marsh, developed the case for the Samples of Anonymised Records (SARs) and conducted an assessment of the risk that an individual or household in the file might be identified. They concluded that the per record risk of identification was negligible but that the research benefit of the SARs was very considerable. Their conclusions were accepted by the Census Offices in 1990 and a research paper was published in the Journal of the Royal Statistical Society, Series A, in 1992.

The SARs were accepted as a statistical abstract under section 4.2 of the 1920 Census Act. The detailed specification to the SARs was subject to independent assessment by Tim Holt - then Professor of Social Statistics at the University of Southampton. The ESRC agreed to commission the SARs from ONS and were given world wide distribution rights. The ESRC then funded the Census Microdata Unit (now CCSR) at the University of Manchester to take responsibility for support and dissemination of the SARs to the UK academic community. In the light of its investment in the SARs, the ESRC also required CCSR to market the SARs to the non-academic sector.

The ESRC and the University of Manchester set up a special licensing system to ensure the safety of the SARs. Each Higher Education Institution was required to sign a licence agreement accepting responsibility for the use of the SARs by their staff and students and nominating an individual - sometimes the Director of Computing Services - as the contact person. Each individual - staff or student - also had to sign a lengthy undertaking in which they agreed not to attempt to identify, or claim they had identified, any individual or household in the SARs and not to pass on the data to an unregistered user. Non-academic institutions signed a 'commercial' licence that imposed similar obligations but which did not require individual employees to sign a document.

The 1991 SARs were released by ONS in the summer of 1993. They are widely seen as one of the success stories of the 1991 Census. There are over 400 academic users across a range of disciplines. CCSR routinely report registration figures to the ESRC and compile a list of publications that is distributed as paper copy and is also on our web site. An indication of the research value of the 1991 SARs can be obtained from our compilation of 'Key Findings', also available from the web site. By any standards it

has been considerable. There has been no known confidentiality breach of the SARs and no known concerns or adverse publicity over confidentiality.

## **2. SARs for 2001: the background**

### *2.1 User consultation*

Whilst the 1991 SARs have been widely used there is, inevitably, room for improvement. Over the last 5 years CCSR has conducted a wide-ranging consultation exercise with users. This has included a survey of all users and also non-users; a considerable number of workshops and meetings aimed at different user groups – for example, local authorities, the commercial sectors, academics; several meetings of the SARs output working group. Users' requests have been compiled and published in the SARs Newsletters, from No. 7 in May 1996 onwards. (Most Newsletters are available from the CCSR web site.)

### *2.2 Assessing the risk of disclosure*

Our programme of work on disclosure control has included a re-assessment of the risk of disclosure from the 1991 SARs as well as an assessment of the change in disclosure risk from the proposed changes for 2001 SARs. This work, reported in Dale and Elliot (2001), showed that the per record risk of the 1991 SARs was rather lower than the assessment by Marsh et al (1992). It was also evident that the risk of identification showed a much greater increase in relation to sample size than to population threshold.

Having established that the 1991 SARs posed even less threat to confidentiality than originally assumed, the parameters of the 2% Individual SAR for 1991 were used as a base-line against which to assess the changes requested for the 2001 SARs. A number of different measures of risk were used, with various scenario-based sets of key variables. The research showed that the two key requirements for the 2001 Individual SAR – an increase in sample size from 2% to 3% and a decrease in the population threshold to about 70K – could be achieved with a risk level similar to that which had been assumed for the 1991 SARs. These two changes to the Individual SAR for 2001 have therefore been requested.

## **3. Proposed sample size and population threshold for 2001 SARs**

*Individual SAR:* 3% sample and reduction in population threshold to 60-70K.

The move to Unitary Authorities will mean that, for 2001 SARs, only 39 per cent of UAs 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 the threshold to 60-70K. The lack of ability to identify separate local authority areas has been a major barrier to use.

In Scotland, lowering the population threshold to 70,000 would allow 27 of the 32 Unitary Authorities in Scotland to be identified. A threshold of 69K would allow the Islands to be separately identified. Since further disaggregation would only be possible with a threshold below 50,000 (well below the range being discussed) we propose that

69,00 represents a sensible target for 2001 SARs in Scotland. A consultation meeting held in Edinburgh in June 2000 provided unanimous support for this proposal.

*Household SAR:* retain same size and structure as 1991 file

#### **4. Proposed geographical areas for 2001 SARs**

*Household SAR:* For the 2001 SAR the Standard Government Office Regions (GORs) should replace the Standard Statistical Regions (SSRs) used in 1991. However, we wish to disaggregate GORs which are more than twice the threshold level. In the 1991 SAR East Anglia was the smallest region with a population of just over 2 million. Disaggregation of larger regions (e.g. Yorkshire and Humberside with a population of 4.7 million) would respect the same population threshold but add extra research value. For Scotland, this would allow a division between lowlands and highlands and for Wales, between rural and South Wales. Dissaggregations within GORs in England would include the important distinction between Inner and Outer London. These distinctions will add considerable value to research and policy analysis.

*Individual SAR:* Where the population size of a Local Authority in the Individual SAR exceeds the agreed threshold by a factor of two or more we recommend subdividing these areas. For example, using a 90,000 threshold, Birmingham, the UK district with the largest population (almost 1 million) could be divided into 10 SAR areas, Leeds into 7 and Sheffield into 5. We propose that disaggregation should be based on wards and CCSR will supply the preferred groupings. This will require further consultation but can only be done after the threshold size has been decided.

#### **5. Proposals for SAR variables**

The detailed specification for each file is enclosed as two documents – one for the Individual SAR and one for the Household SAR. This aims to list the detail required for each variable and also includes new variables not in the 1991 SARs.

The document also highlights any changes required in variable detail between 1991 and 2001. It represents a first draft for discussion with ONS. There will, inevitably, be some additional, minor amendments made to this specification during the next 12 months in response to discussions with ONS and further advice from users.

The following annotations have been used:

- \* questions new to the 2001 Census
- \*\* an extension of categories by comparison with 1991
- \*\*\* additional variables, present in 1991 but not included in dataset

*Documents:*

Statistical specification for Individual SAR

Statistical specification for Household SAR

There are a number of specific issues which are discussed in detail in the following sections.

## 6. Migration information in the SARs

In 1991 the Individual and Household SAR provided only the region of former usual residence for an individual who had moved in the previous 12 months and a banded variable giving 13 categories of distance moved. The Region variable identified Scotland and Wales as separate regions. Northern Ireland was included with the final category 'Outside GB'. The SARs for Northern Ireland used the same categories, except that Inner and Outer London were combined as London. It was thus not possible to distinguish moves from the Irish republic to Northern Ireland.

The lack of information on the area of origin of migrants rendered the 1991 SARs of little value to migration analysts. There is therefore considerable concern amongst migration experts in the UK that the 2001 SARs should provide better information. Extensive consultation has been carried out by Paul Boyle and Tony Champion, including a survey and a meeting of experts. The conclusions are set out below:

### 6.1 Information on area of origin (MIGORGN): order of preference:

#### *Individual file:*

1. The SAR area of residence in 2000 should be coded using the same level of detail as the 2001 area of residence. This would allow users to re-constitute areas to regions in order to achieve continuity with 1991 SARs and would be sufficiently detailed to allow distance to be derived using centroids.
2. The second preference to SAR area is to provide groupings of SAR areas but below 'virtual' county level
3. Groupings of SAR areas at county level
4. Regions – or sub-regions – as used in the 2001 Household SAR for area of current residence

#### *Household file:*

1. Area of origin based on same spatial units as Household SAR
2. Area of origin based on Government Office Region

### 6.2 Information on distance moved: order of preference

#### *Individual and Household SAR*

1. Actual distance moved, rounded to nearest km or mile
2. Distance grouped with two additional bands at lower end and two additional bands at upper end as shown below and in detailed variable specification.

-99	Not applicable
-1	From outside GB
1	0-2 km
2	3-4 km
3	5-6 km

4	7-9 km
5	10-14 km
6	15-19 km
7	20-29 km
8	30-39 km
9	40-49 km
10	50-59 km
11	60-79 km
12	80-99 km
13	100-119 km
14	120-149 km
15	150-199 km
16	200-249 km
17	250 km and over

### 6.3 Type of move made (MIGIND)

An additional variable (MIGIND) has been requested which summarises the type of move made. This is particularly helpful for local authorities as it allows them to establish movement into their authority. The categories are given below:

#### *Individual and Household file*

-99	Not applicable
0	Same address
1	Move within a SAR/LAD area
2	Move between SAR/LAD area but within county
3	Move between counties within a region
4	Move between regions but within country
5	Move between countries but within UK
6	Move from outside UK

### 6.4 Household migration indicator

We propose that 'household migration indicator', from the 2001 census database, should replace the 1991 Household file variable 'wholly moving household. The six categories are:

#### **MIGHUK – HOUSEHOLD MIGRATION INDICATOR**

-99	Not applicable
1	Whole household lived at same address
Wholly moving households	
2	Lived elsewhere one year ago, within same area
3	No usual address one year ago
Inflow	
4	Lived elsewhere outside area but within district (or UA)
5	Lived elsewhere outside district (or UA) but within UK
6	Lived elsewhere outside UK
Outflow	
7	Moved out of area but within district (or UA)
8	Moved out of district (or UA) but within UK
9	Partly moving household

The exact information on the SAR database needs to be the subject of detailed consultation within ONS and with key experts.

## **7. Occupational coding**

The SAR user community is very concerned about the reduction in the population for whom there will be information on occupation, industry and social class. The requirement is to find a way to code 16-64s who last worked more than 5 years ago but less than 10, and 65-74s not currently working who worked in the previous 10 years.

ONS has made a commitment that this coding will be added to the SARs although the means of doing this has not yet been agreed. The first essential step is to ensure that the string text recorded for members of the SARs sample is extracted from the census database. This will provide a basis for classification.

### *7.1 The size of the problem*

In the Individual SAR, the numbers of respondents aged 16-64, not currently working but who had worked within the last 10 years on the 1991 SARs was: 128,000

Assuming that 70% of these worked within the last 5 years, we are left with 30% (38,400) who worked 5-10 years ago. Amplified for a 3% sample this is: 54,600. Of those aged 65-74, 35,400 were not currently working but had held a job in the last 10 years. With a 3% sample this would be 53,100.

Repeating these calculations for the 1% Household SAR, we have 61,500 individuals aged 16-64 not currently working but who had a job in the last 10 years. Again, assuming 70% worked within the last five years we have  $61,500 \times 30 = 18,450$  to be coded. Of those aged 65-74, 17,673 are not currently working but held a job in the last 10 years.

Thus the estimated total number of individuals who have occupational information which will not be coded is:

73,800 on a 2% Individual SAR  
107,700 on a 3% Individual SAR  
36,000 on the 1% Household SAR

Urgent discussions are needed to agree strategies for getting this information coded. Possibilities include:

1. Coding by ONS using the Lockheed Martin software – not feasible.
2. Coding by ONS using their occupational coding software – CASOC.
3. Coding within ONS but conducted by an academic using either software packages
4. Setting up a methodological project to establish the consistency of coding based on the Lockheed Martin software by comparison with CASOC. If this were done using the SARs then we would be able to re-code some of the respondents coded by Lockheed Martin and identify whether the two methods produced any significant differences. Once the process was in place and working we could then extend the occupational coding to some or all of the SARs sample who would not be coded as part of the main census coding. We would be able to seek the marginal cost of this additional coding from ESRC, although, realistically we could not ask for a large sum.

If this additional coding is not done the research value of the SARs will be seriously impaired. The reasons for this are set out below:

### *7.2 The need for full coding of occupation and industry on the SARs*

The 2001 Census schedule asks for occupation and work-place information from all respondents aged 16 and over and under 75 who have ever worked. However, for financial reasons ONS have decided that they are only able to code occupation, industry and socio-economic class for respondents :

Aged 16-65 who have worked in last 5 years; or 65-74 and currently working

This means that, although information has been collected, it will not be coded for:

16-65s who last worked more than 5 years ago

65-74s not currently working

This has important implications for the research value of the SARs. We know from analysis of the 1991 SARs that restricting occupation to those with a last occupation 10 years ago leads to serious bias. For example, about 25% of households, nationally, had a head of household with no recorded occupation and thus no social class. In Liverpool this went up to 37%. If coding is further restricted in the 2001 Census to those who held a job in the previous five years these figures will be considerably higher.

The percentage of those without an occupation is highest for older people, for women, students, the unemployed and those on Government schemes. These groups will be disproportionately omitted from any analyses that require a measure of social class. Social class based on last occupation continues to be of considerable relevance in predicting outcomes such as ill health. With no income question in the census there are few alternative measures of life-style and life-chances. Table 1 shows the profile of those who recorded no occupation in the previous 10 years in the 1991 Census. We expect that, if occupational coding is further restricted, these percentages will be considerably higher – particularly for those aged 65–74 who will only have a coded occupation if they are currently working.

Whilst the 2001 Census will provide a new question which allows us to calculate how long ago someone last worked, we will not be able to relate to this the job they last held – unless it happens to be less than five years ago and they are under 65!

Table 1: Individuals aged 16 and over, 2% GB SAR

Variables	% with No RG Class*
Sex: Male	18
Female	36
Age: 16-24	26
Age: 25-44	10
Age: 45-64	27
Age: 65 and over	75
Economic activity:	
Employed	0*
Government Scheme	26
Unemployed	27
Student	62
Permanently sick	53
Retired	68
Other inactive	64
Family type:Single,'no family'	46
Married, no children	32
Married, dep.children	16
Married, non-dep. Children	16
Lone parent, depl child	35
Lone parent, non-dep. Children	30
All adults, 16+	27.6

\*This refers to people who were not asked for an occupation in the 1991 census. A further 1.4% of people with a job in the last 10 years failed to give a usable response.

## 8. Sampling design for SARs 2001

As far as possible we wish to retain the sampling design used for the 1991 SARs. However, there are two key differences between the 1991 and 2001 Census that will affect the sampling design:

1. The 2001 Census will 100% code all data. Therefore the SARs can be drawn from the 100% database whereas, in 1991, they were drawn from the 10% sample in order to provide full coding on all topics. This should lead to a slight improvement in the accuracy of the Individual sample by comparison with 1991.
2. The 1991 SARs omitted 'wholly imputed households' which contained information only for the 100% coded topics and therefore had no information on 'hard to code' items such as occupation, industry and relationship. In 2001 the One Number Census will produce a complete database 100% coded for all individuals and households. All census outputs will be drawn from the database. Therefore, unlike 1991, the SARs will contain individuals and households which have been imputed as part on the One Number Census. As discussed below, we would like these individuals and households to be flagged in the SAR data files. The inclusion of imputed individuals and households should provide greater consistency between the SARs and the Standard Area Statistics.

3. We assume that any record-swapping will be done before the SARs sample is drawn. This will ensure that the SARs cannot be used to unpick record swapping that crosses SAR areas.

A summary of the 1991 sampling design, extracted from Campbell et al, (1996) is set out below and a technical document recording the output processing used by ONS for the 1991 SARs is enclosed.

The coding of the 1991 Census was divided into two stages. Easy to code information, such as sex, date of birth, marital status and country of birth, was processed first for all forms. A 10 per cent sample of these partially coded forms was selected and the remaining hard to code questions, mainly those relating to occupation, industry and qualification, were coded. This 10 per cent sample was then used to extract the SARs. The 10 per cent sample was selected during processing from the 100 per cent records, using three levels of stratification: first each county was treated separately; second, each processing unit of 50 consecutive enumeration districts was treated separately; third, each processing unit was split into strata consisting of 10 households or 10 persons in communal establishments. The 1991 census imputed records for 'wholly absent' households, but these households were not included in the 10 per cent sample, and were therefore excluded from the SARs. The 10 per cent sample, therefore, consisted of one household selected at random from each stratum of 10 consecutively recorded households, and a similar sample of persons in communal establishments. Within blocks of 50 EDs, the strata ran continuously from the first household in the first ED to the last household in the 50th ED and the first person in the communal establishment in the first ED to the last person in the last communal establishment in the 50th ED.

The sample design for the SARs was divided in two stages, with the one per cent Household file selected first. All fully-coded household forms were ordered geographically by county and enumeration district in England and Wales and by region and output area in Scotland. They were then grouped into batches of 10, and one household selected at random from each batch. All sampled records were then scrambled to prevent geographical tracing, before being released.

The two per cent Individual sample was then taken *from the remaining households*, hence there is no overlap between the two samples. Individuals in the remaining households were stratified into groups of nine, and two individuals selected from each group at random. It is therefore possible that more than one individual may be selected from one particular household. For example in the following sequence of households:

household 1: 1 person  
household 2: 3 persons  
household 3: 2 persons  
household 4: 1 person  
household 5: 4 persons  
household 6: 1 person  
household 7: 3 persons  
household 8: 2 persons  
household 9: 2 persons

The individuals would be grouped into groups of 9 as follows:  
(number refers to household number)

1 2 2 2 3 3 4 5 5 | 5 5 6 7 7 7 8 8 9 | 9

The first group of nine individuals runs up to and includes the second person in household 5; The second group of nine extends to the first person in household 9. Two individuals would then be selected from each group, with the possibility of all four members of household 5 being selected, two from each group.

For the final stage of the sample design, individuals in communal establishments were stratified into groups of five, and one individual selected at random from each group. Once again, the records were scrambled before being released to prevent the geographical tracing within a SAR area.

#### *Amendment to sampling design for 2001 SARs*

Create strata of 100 households and select 1 out of 100 for the Household SAR. Then select 2 (3) out of 100 individuals from the remaining 99 households.

#### *8.1 Calculation of sampling variance and design factors in the SARs*

Some time after the release of the 1991 SARs Dr Malcolm Campbell from MIMAS, University of Manchester conducted a programme of work to calculate sampling variance and design factors for all the variables in both the Individual and Household SARs. We need to discuss with ONS the feasibility of doing these calculations at the time of extraction of the SARs, rather than as an additional operation some time later.

### **9. Recording imputation in the SARs**

As part of the One Number Census individuals and households will be imputed so that the 2001 Census database will represent a complete population count, corrected for under-enumeration. The SARs will be drawn from this database.

#### *8.1 Wholly Imputed individuals and households*

For some SAR users it will be important to be able to identify imputed individuals or households and for this reason ONS have agreed to flag the records for individuals that have been imputed. Users wish to have a variable attached to each individual in the Individual and Household files to signify the following:

- Not imputed individual (ie minimum of 4 core items recorded on schedule)
- Imputed individual in non-imputed household
- Imputed individual in imputed household

#### *8.2 Item imputation for individuals and households*

Missing or out of range information in the 2001 Census will be 'filled-in' by a method of donor imputation. Initial processing suggests that about 30% of records have one or more items imputed - considerably more than the 1991 Census. Users have requested a flag or marker to indicate item imputation. We would like the following flags added to the SARs database:

- 1 A variable recording the number of imputed items for a given individual or household
- 2 A variable recording the level of imputation required for the record
- 3 A flag (eg a minus sign or other character) preceding any imputed value.

For most purposes this flag (3) would be ignored - and would not be read into the main SPSS or SAS files distributed. However, it will provide a valuable basis for analysis of which respondents found the census hard to answer and which questions caused most problems. This information has significant research value particularly in understanding patterns of response to the census.

### **10. The addition of an area level classification to the SARs**

There is a strong requirement for the addition of an area-level classification to the 2001 SARs. The classifications added to the 1991 SARs have demonstrated the research value of the variable, particularly in the ability to add a locality dimension to analysis. Several high quality papers have been published which use the area-level classification in the 1991 SARs.

In 1991 the ONS ward-level classification was added to the Household file and GB Profiles, an ED-level classification developed by Stan Openshaw, was added to the Individual file.

For 2001 SARs we require:

An area-level classification to be added to both SARs

We need the same classification for both files and it should relate to Census Output Areas (ie about 250 people).

The classification should be hierarchical - ie it should be possible to collapse it

The 1991 confidentiality criteria applied by ONS were:

- on the Individual SAR there had to be at least 10 EDs in the same category within any SAR area and
- on the Household SAR there had to be at least 5 wards in the same ward-type within any region.

There may, of course, be no EDs or wards represented in a category.

We anticipate that, for 2001, ONS will want to ensure that there should be no direct look-up table available to provide the users with the list of OAs or wards that relate to each category of the classification as this would provide significant additional geographical information about location. This therefore means that a ready-made classification cannot be used and we need to ask a possible supplier to produce a tailor-made classification.

We do not yet have a classification available. We may need to derive a bespoke classification. Decisions have yet to be made as to whether this should contain solely data from the 2001 Census or additional data from Neighbourhood Statistics.

As a classification would need to be based on the 2001 Census output areas - and would probably include some 2001 Census statistics - it would not be available at the time the SARs were produced but would need to be added to the database later on.

We therefore need in principle agreement from ONS that the addition of a classification is acceptable and agreement to add it to the 2001 SARs about 6-12 months after the first release of the data.

The cost of adding the variable to each file this should be included as a separate item in the figure quoted to the ESRC.

## **11. Small Area Microdata**

The case for small area microdata was set out in a report sent to ONS in January 2001 and a rather shorter paper which was submitted to the Journal of the Royal Statistical Society at the same time. A report on the project was also submitted to the ESRC and was the subject of extensive academic review. It was graded 'outstanding'.

The ESRC has allocated funding to enable the purchase of small area microdata. However, we want ONS to consider this request separately from that for the Individual and Household SARs and after agreement has been reached for the SAR files. In other words, we do not want the decisions over the two SAR files to be influenced by the request for an additional Small Area Microdata file.

## **12. Arrangements for purchase and dissemination of 2001 SARs**

The ESRC and the Joint Infrastructure Funding Council (JISC) have agreed to purchase samples of microdata for use by the academic sector. CCSR have been awarded a 5-year contract, from 2001-2006 to support and disseminate samples of microdata.

For all 2001 Census products there will be a common registration procedure conducted by the Data Archive. However, we envisage that the safeguards to protect the confidentiality of the 2001 SARs will be similar to those for the 1991 SARs.

The ESRC/JISC require a final costing from ONS and the supply of an order form. They expect to be invoiced once the data have been accepted and that payment will be required within 30 days of acceptance.

Costs should be broken down to distinguish the supply of the Individual and Household files; the cost of additional coding for occupations and industry; and the addition of an area level classification for both files. The costs of a small area microdata file should be shown separately.



## Appendix

### Supporting documents

Brown, M. and A. Dale. (1998), "A survey of SAR users, their requirements for 2001 SARs and their view on dissemination and support", CCSR Working Paper 6, downloadable from: <http://les1.man.ac.uk/ccsr/publications/working/saruser.htm>

Dale, A. and Elliot, M. (2001) Disclosure assessment of 1991 SARs and proposals for 2001 SARs in *JRSS(A)*, Vol.164, No.3, and <http://les1.man.ac.uk/ccsr/publications/>

Statistical specification for Individual SAR

Statistical specification for Household SAR

ONS SAR sample specification for 1991

Dale, Tranmer, Elliot, Martin, Brown, Pickles and. Fieldhouse, *Report to ONS on 'The case for small area microdata from the census of population'*, CCSR, University of Manchester, January 2001

Campbell, M., C. Holdsworth, T. Payne, and A. Dale (1996), *Sampling Variance and Design Factors in the Samples of Anonymised Records*, CCSR Occasional Paper No. 6, Manchester: CCSR, University of Manchester.