Samples of Anonymised Records from the 1991 Census for Great Britain

by Angela Dale[1]
Census Microdata Unit,
University of Manchester

Introduction

For the first time in a British census, the 1991 statistical output included Samples of Anonymised Records (SARs). Known as Census Microdata or Public Use Sample Tapes in other countries, SARs differ from traditional census output of tables of aggregated information in that abstracts of individual records are released. The released records do not conflict with the confidentiality assurances given when collecting census information since they contain neither names or addresses nor any other direct information which would lead to the identification of an individual or household. Essentially three per cent of records have been released in two samples. The SARs offer users the freedom to import individual-level census records into their own computing environment and the ability to produce their own tables or run analyses which are not possible using aggregated statistics.

Background to the release of the SARs

Requests had been made for SARs to be released from previous censuses in Great Britain. The principal stumbling block in the past had been an argument as to whether SARs could be considered a statistical abstract for release under Section 4.2 of the Census Act 1920 at the request and expense of user(s). Furthermore, in the past, requests for SARs had failed to reach a compromise between those (often geographers) wanting fine grain areal detail and those (often sociologists and demographers) wanting fine grain detail on other variables such as occupation.

The 1991 Census White Paper (Her Majesty's Government 1988), however, announced:

> "The Government intends that results from the 1991 Census should wherever practicable be made available in a convenient form to meet users' needs"

Legal advice having been received that SARs could be deemed statistical abstracts, the White Paper went on to say:

> "Requests for abstracts in the form of samples of anonymised records for individual people and households ... would also be considered, subject to the overriding need to ensure the confidentiality of individual data".

The Economic and Social Research Council (ESRC) set up a working party to negotiate with the Census Offices and present a formal request. Their report, presented to the Census Offices in 1989 (subsequently published as Marsh, Skinner et al. 1991) concentrated on the benefits of releasing SARs, the uses to which they would be put, and also an assessment of the confidentiality risks involved in releasing SARs.

The request was mentioned by Ministers during the debate on the Census Order in Parliament at the end of 1989. Having considered the request, the Registrars General for England and Wales and for Scotland announced in July 1990 that they had agreed in principle to the release of SARs from the 1991 Census. There then followed detailed work by the Census Offices and ESRC in developing the statistical specification. An independent technical assessor, Professor Holt (University of Southampton), was appointed to advise the Registrars General on the confidentiality aspects and to write a report to Ministers. Following receipt of the report it was announced in March 1992 that two SARs from the censuses in England and Wales and in Scotland would be produced and released to ESRC. Similar SARs for Northern Ireland have also been made through an ESRC purchase. These allow the production of harmonised SARs for the whole of the United Kingdom.

Details of the SARs

Two SARs have been extracted from the GB censuses:

1 a two per cent sample of individuals in households and communal establishments; and
2 a one per cent hierarchical sample of households and individuals in those households.

The two per cent SAR has finer geographical detail and the one per cent SAR has finer detail on other variables, thus providing a solution to the conflict between users' demands discussed above.

The two per cent individual SAR contains some 1.12 million individual records (1 in 50 sample of the whole population enumerated in the census). It was selected from the base which lists persons at their place of enumeration. Details are given as to whether or not the person was a usual resident of that household, and if so (and enumerated in a household) whether they were present or absent on census night. The following other information is given for each sampled individual:

- details about the individual ranging from their age and sex to their employment status, occupation and social class;

- details about the accommodation in which the person is enumerated (such as the availability of a bath/shower and the tenure of the accommodation) or, if they were in a communal establishment, the establishment type (hotel, hospital, etc.);

-information about the sex, economic position (in employment, unemployed, etc.), and social class of the individual's family head; and

-limited information about other members of the individual's household (such as the number of persons with long-term illness and numbers of pensioners).

In effect, all the census topic variables listed are on the file; the only exceptions are variables either suppressed or grouped to maintain the confidentiality of the data. In all, there are about forty pieces of information about each individual, and the size of the raw data file, before any new variables have been derived and before any data compression techniques have been applied, is around 80 megabytes.

The one per cent household SAR contains some 240,000 household records together with sub-records, one for each person in the selected household. Information is available about the household's accommodation together with information (similar to the two per cent sample) about each individual in the household and how they are related to the head of the household. The raw data is supplied as a hierarchical file in non-software specific character format (one line of information about housing and household, followed by one line of information about each individual in the household).

The full details of the information provided in both SARs are given in the Codebook and Glossary files produced by the Census Microdata Unit. Table 1, however, provides summaries by describing the information collected on the census form, the detail of coding of that information on the census database, and in how much detail that information is being released in the SARs.

The sampling procedure used

Census data goes through two separate coding processes. The easy to code information such as housing details, sex, date of birth, and country of birth is processed for all forms (100 per cent). The harder to code information such as occupation and industry is only processed for 10 per cent of forms. Both SARs were drawn from the 10 per cent sample so that they contain information from the whole of the census form. A detailed description of the sampling scheme for the SARs is given in Dale and Marsh (1993, chapter 11).

Confidentiality protection in the SARs

The census offices in some European countries have refused to release microdata because they believe, on the basis of research such as that conducted by Paass (1988) and Bethlehem et al. 1990), that the risks of disclosing information about respondents' identities are too high. Much of this work is concerned with how many people have unique combinations of census characteristics which would make them open to identification. The Economic and Social Research Council Working Party which negotiated the release of the SARs took the view that uniqueness was only one part of a four-stage process of disclosure: data in the microdata file would have to be recorded in a compatible way to that in an outside file, the individual in an outside file would have to turn up in a SAR, the individual would have to have unique values of a set of key census variables and the matcher would need to be able to verify this uniqueness. Rough estimates of the size of risk at each stage were made; when cumulated, the risks of disclosure appeared very low; multiplying the various probabilities together, the working party concluded that the risk of anyone in the population being identifiable from their SAR record were extremely remote; their best estimate was something of the order of 1 in 4 million. (For more details of such calculations, consult Marsh, Skinner et al. 1991, Marsh, Dale and Skinner (1994) and Skinner, Marsh et al. 1992.) The arguments put forward were important in persuading the census offices to release the SARs suitably modified to protect anonymity where this was felt at risk. In this section the various disclosure protection measures taken are described.

Sampling as protection

The low sampling fractions of the SARs offer a strong source of disclosure protection for sensitive data. It not only reduces the actual risk that a particular individual can be found in the census output, but it probably has its greatest effect by reducing the chances that anyone would make the attempt at identification by this means. The two SARs (a one per cent sample of households and a two per cent sample of individuals) are sufficiently small to offer a great deal of protection; the samples do not overlap so that the detailed household or occupational information available on the household file cannot be matched with the detailed geographical information available on the individual file.

Restricting geographical information

One of the key considerations which may affect the possibility of disclosure of information about an identifiable individual or household is the geographical level to be released (i.e how much detail is given about where the person was enumerated). The full census database holds information at enumeration district level (about 200 households or 500 persons in each ED) and even at unit postcode level (about 15 households). If released, such detailed geography would obviously pose a confidentiality risk. Empirical work and comparisons with SARs released in other countries showed that a sensible level for release would be areas equivalent to large local authority districts for the individual (2%) SAR.

To be separately identifiable, the decision was taken that an area had to have a population size of at least 120,000 in the mid-1989 estimates. The primary units used were local districts; only one geographical scheme was permitted, or smaller areas could be identified in the overlap, say between a local district and a health district. A population size of 120,000 is slightly higher than the lowest level of geography permitted in the US SARs (100,000), but it still has the advantage of allowing all non-metropolitan counties in England and Wales, most Scottish regions, all London boroughs (except the City of London), and all metropolitan districts to be separately identified.

Smaller local authority districts (under 120,000 population) were grouped to form areas over 120,000. Several rules were used to decide how districts should be amalgamated where this was necessary. First, the integrity of county/Scottish region geography was always maintained, where possible. Secondly, districts which achieved the minimum population threshold on their own were left intact, where possible; and smaller areas were grouped with each other. Thirdly, grouping was done on the basis of contiguity. And finally, if there was a choice left once the above criteria had been met, areas were grouped on the basis of their apparent social and historical similarity.

The one per cent household SAR, because of its hierarchical nature (i.e. statistics about the household and all its members), is more of a disclosure risk. For this reason it was decided that, for this SAR, the lowest geographical detail revealed would be the Registrar General’s Standard Regions, plus Wales and Scotland. The only exception is that the South East is split into Inner London, Outer London, and the Rest of the South East Region.

It should be noted that the order of records in both SARs has been re-arranged before the Census Offices release them. This is to prevent any possible tracing of individuals or households back through a region or district.

Suppression of data and grouping of categories

Some alterations have been made to the data to reduce the number of rare and possibly unique cases. The extent to which the variables on the local base have been either suppressed entirely or modified by grouping small categories before release in SARs is shown in Table 1.

Information which is unique in itself, such as names and addresses, has been omitted altogether; (technically these variables have not been suppressed since they are never put on the computer). Precise day and month of birth have been suppressed.

The thresholding rule

The degree of detail permitted on other variables was the subject of a thresholding rule which ensured that the expected value of any category at the lowest level of geography on any file was at least 1. The threshold, when operationalised, dictated that a category must have 25,000 cases in it in the GB file before it could be released on the individual SAR, or 2,700 cases before it could be released on the household SAR.

With some other variables, the smaller categories have been grouped, either across the entire range of the variable or only at the extremes (a process know as “top coding”). The rule used to decide the level of detail to be released was to group information categories to a sufficient detail so that, on average, the expected sample count would be at least one for each category of each piece of information for the lowest geographical area permitted on each SAR.

Some justification for restricting attention to the distribution of the univariate categories of each variable in turn was given by Marsh et al (1994). They demonstrated that the risk of an individual having a unique combination of values of a set of variables could be predicted with a high degree of certainty simply from knowledge of their membership of rare categories of each variable taken singly. The precise cut-off at an expected value of 1 was set at a value sufficiently high to give reasonable protection of anonymity.

The rule was applied to each census variable. Expected counts were obtained by using 1981 Census frequency counts (supplemented by more recent surveys, for example the Labour Force Survey) at the national level for the whole population. To obtain expected counts, the count of 1 per category per SAR area was grossed up to the national level:

$$C = 1/X * (Y/Z)$$

where

C = expected count at the national level
X = sampling fraction (1/50 for individual SAR and 1/100 for household SAR)
Y = national population (56 million)
Z = smallest geographical area population (120,000 for individual SAR and 2.1 million (East Anglia) for household SAR

Thus 25,000 and 2,700 were the two thresholds used for the individual and household SARs respectively. In theory, a small amount of random noise could have been added to certain variables in a manner analogous to the procedure adopted for the small area statistics. A technique similar to this has been used in the 1990 US Census for example: geography has been subject to a degree of perturbation by switching a small number of similar households between nearby areas (Navarro et al. 1990). However, the natural levels of noise in the data, combined with the analytical difficulties of minimising bias to both measures of location and spread by such techniques in a multipurpose file led to perturbation not being implemented in any form for the SARs.

Grouping of variables

When expected frequency counts fell below the threshold, categories were grouped. With some variables, grouping was only required at one end of the distribution: thus rooms were top-coded above 14 and the number of persons in the household was top-coded above 12. Two variables were both grouped and top coded; with age, 91 and 92 were grouped, 93 and 94 were grouped and 95 and over was top-coded; with hours of work, 71-80 hours per week has been grouped and the rest top-coded above 81.

When variables were not measured on a numeric scale, judgments had to be made about which categories to put together. Classifications for census data are often hierarchical. For example, for the Standard Occupational Classification there are 371 unit groups, 77 minor groups, 22 sub-major groups, and 9 major groups. In cases such as these, small categories could be amalgamated to the next level in the hierarchy. In other cases, detailed advice was sought from subject experts about how the groups should be formed.

In the case of three variables in the two per cent individual SAR, it was deemed necessary to further group categories, even though they contained numbers which fell above the threshold: occupation, industry, and subject of qualification. As a result of advice received from the Technical Assessor, occupation was reduced from the 220 categories proposed (out of a possible 371) to 73; similarly industry was cut from a possible 334 to 60 and subject of educational qualification from a possible 108 to 35. (Almost full occupational detail remains on the one per cent household SAR, however.)

There were other factors which determined the detail to be released:

- Categories of occupations and industries in the public eye were grouped further than mathematically necessary to guard against disclosure; for example, actors/actresses and professional sportsmen/women;

- Large households were seen as a disclosure risk in the household sample. Applying the frequency rule to size of household, a large household in the 1981 Census was estimated to be one of 12 persons or more. Consequently, only housing information is given for households containing 12 or more persons. No information about the individuals in the household is given.

Table 1

Details of the information in the two Samples of Anonymised Records from the 1991 Census of Great Britain

Item	Household (1%) sample		Individual (2%) sample	
	No. of categories (maximum*)	Other details	No. of categories (maximum*)	Other details
Geographical area of renumeration	12	Standard regions of England (with split of South East into	278	Local authority districts over 120,000 population. Others
			Inner London, Outer London	
amalgamated to form areas over and Rest), Wales and Scotland	120,000	Housing/household information		
Accommodation type	14 (14)	Detached, semi-detached or	As household sample	
			terraced house; purpose built	
		flat in a commercial or residential building; converted or not self-contained accommodation in a shared house or flat		
Availability of amenities				
ù bath/shower	3 (3)	Exclusive, shared or no use	As household sample	
ù inside WC	3 (3)	Exclusive, shared or no use	As household sample	
ù central heating	3 (3)	Full, part or none	As household sample	
Cars (number of)	4 (4)	0, 1, 2, 3 or more	As household sample	
Floor level (lowest), of accommodation (Scotland only)	7 (101)	Basement, ground, 1st/2nd, 3rd/4th, 5th/6th, 7th to 9th 10th or higher	As household sample	
Number of household (accommodation) spaces in dwelling	4 (35)	Top coded: 4 or more	Not included	
Number of persons (enumerated) in household	12 (99)	Top coded: 12 or more	Not included	
Number of residents in household		Derivable	4 (99)	0, 1, 2 to 5, 6 or more
Number of dependent children in household		Derivable	2 (99)	0, 1 or more
Number of pensioners in household		Derivable	2 (99)	0, 1 or more
Number of persons with long-term illness in household		Derivable	2 (99)	0, 1 or more
Number of persons in employment in household		Derivable	3 (99)	Top coded: 2 or more
Number of rooms	15 (19)	Top coded: 15 or more	Not included	

Number of persons per room		Derivable	5	Ranging from less than 0.5 to more than 1.5
Tenure	10 (10)	Owner occupier or rented (public sector or private)	As household sample	
Wholly moving household indicator	2 (2)	Yes (all resident household members are migrants from the same address) or No	Not included	
Individual information				
Age	94 (111)	Single years 0 to 90, 91/92, 93/94, 95 and over	As household sample	
Status in communal establishment		Not applicable	3 (4)	Visitor, resident staff or resident non-staff
Type of communal establishment		Not applicable	15 (35)	Hotal, hospital, nursing home etc.
Country of birth	42 (102)		As household sample	
Migrants _ distance of move (km)	13	5, 10, 20 and 50 km bands; top coded above 200 km	As household sample	
Distance to work (km)	8	10 km bands; top coded above 40 km; 0_9 km band split 0-2, 3-4 and 5-9	As household sample	
Economic position				
primary	10 (12)	Employee, self-employed, unemployed, student, retired etc.	As household sample	
secondary	7 (10)		As household sample	
Economic position of family head		Derivable	3 (12)	Employed, unemployed or inactive
Ethnic group	10 (10)		As household sample	
Family head indicator	2 (2)	Yes or no	Not included	
Family number	5 (5)	Used to identify individual's family	Not included	
Family type	8 (8)	Married or cohabiting couple family with or without children or lone-parent family	As household sample	
Gaelic language (Scotland only)	5 (8)	Ability to speak, read or write Gaelic	As household sample	
Hours worked weekly	72 (99)	Single hours 0_70, 71 to 80, 81 or more	As household sample	
Industry of employees and self-employed	185 (334)	Mainly third digit (groups) of 1980 SIC	60 (334)	Mainly second digit (classes) of 1980 SIC

Limiting long-term illness	2 (2)	Yes (individual has illness) or no	As household sample	
Marital status	5 (5)		As household sample	
Migrant - geographical area of former residence	13	Standard regions of England (with split of South East), Wales, Scotland, outside GB	As household sample	
Occupation	358 (371)	Mainly unit groups of	73 (371) 1990 SOC	Mainly minor groups
of 1990 SOC				
Number of higher educational qualifications	3 (7)	0, 1, 2 or more	As household sample	
Level of highest qualification	3 (3)	Higher degree, first degree, above GCE A-level	As household sample	
Subject of highest qualification Classification	88 (108)	Mainly third digit of Standard Subject Classification	35 (108)	Mainly second digit of Standard Subject
Relationship to household head	17 (17)		8 (17)	
Resident status	3 (3)	Present resident, absent, resident, visitor	As household sample	
Sex	2 (2)		As household sample	
Sex of family head		Derivable	2 (2)	
Social class	8 (8)		As household sample	
Social class of family head		Derivable	8 (8)	
Socioeconomic group	19 (20)		As household sample	
Term-time address of students and school children	4	Inside or outside region of usual residence	As household sample	
Transport to work (mode)	10 (10)		As household sample	
Visitor _ geographical area of residence	13	Standard regions of England (with split of South East), Wales, Scotland, outside GB	As household sample	
Welsh language (Wales only)	5 (8)	Active use of (speak, read or write)	As household sample	
Workplace	5	Inside or outside region of usual residence	5	Inside or outside SAR area of usual residence

* The maximum number of categories as available on the full census database.

given.

- Geographical information for such items as workplace and migration (address one year before census) has been heavily grouped. This is because of the high likelihood of uniqueness of such information when used in conjunction with area of residence.

Dissemination

The licensing and distribution of the SARs is the responsibility of Manchester University who have a contract with the ESRC. The SARs may be used for both academic and non-academic purposes. All Higher Education Institutions (HEI) are required to sign an End User Licence Agreement which makes the HEI responsible for those members of their institution who are using the data. Users within each institution must be either members of staff or students and must sign a further individual registration form which contains a binding undertaking to respect the confidentiality o the data. Specifically, users have to guarantee not to use the SARs to attempt to obtain or derive information about an identified individual or household, nor to claim to have obtained such information. Furthermore, they have to undertake not to pass on copies of the raw data to unregistered users, and the Census Microdata Unit has the responsibility of auditing their use of the data. They must sign a statement that they understand that the consequences of any breach of the regulations on the part of any user in a specific institution can lead to the withdrawal of all copies of the data from that institution. Non-academic organisations sign a similar End User Licence Agreement and undertake not to allow the data to be user other than by their employees.

The data is free for the purposes of academic research; to get the data free the researcher must be doing the research in an institution qualified to receive an ESRC award, and the research must be funded either by the Universities Funding Council or one of the Research Councils. When the data is used either by those outside the academic sector or by researchers in universities for sponsored research, a charge is made for the data. In order to encourage a high volume of usage of a product whose advantages may not yet be well appreciated in Britain, these charges are being kept extremely low; an entire national SAR can be bought for £1,000 + VAT, and subsets of a county or local district for £500.

[1] Paper presented at IASSIST 21st Annual Conference May 9-12, 1995, Quebec City, Canada.

References

Barnett, V. (1991) *Sample Survey Principles and Methods*, London: Edward Arnold

Bethlehem J G, Keller W G and Pannekoek J. (1990) Disclosure control of microdata, *Journal of the American Statistical Association*, 85: 38-45

Breton, R, Isajiw, W, Kalback, W and Reitz, J (1990) *Ethnic Identity and Equalit*y, Toronto: University of Toronto Press

Goldstein, H (1987) *Multilevel Models in Educational and Social Research*, London: Charles Griffin and Company

Her Majesty’s Government (1988) *White Paper (Cm 430), 1991 Census of Population*, HMSO

Li, P (1999) *Ethnic Inequality in a Class Society>*, Toronto: Wall and Thompson

Marsh C, Skinner C, Arber S, Penhale B, Openshaw S, Hobcraft J, Lievesley D and Walford N. (1991) The case for samples of anonymised records from the 1991 Census, *Journal of the Royal Statistical Society* (A), Vol 154 (2): pp 305-340

Marsh, C, Dale, A and Skinner, C (1994) Safe data versus safe settings: access to microdata from the British Census, *International Statistical Review*, 62,1, 35-53

Paass G. (1988) Disclosure risk and disclosure avoidance for microdata, *Journal of Business and Economic Statistics*, 6(4): 487-500

Skinner,C.J,Holt,D. and Smith T.M.F. (Eds)(1989) *Analysis of Complex Surveys*, New York: Wiley

Skinner, C, Marsh, C, Openshaw, S and Wymer, C (1992) *Disclosure control for census microdata*, University of Southampton, mimeo.

Wolter,K.M.(1985) *Introduction to Variance Estimation*, New York: Springer Verlag