

2. What are the Samples of Anonymised Records?

2.1 Introduction

The Samples of Anonymised Records are samples drawn from the 1991 and 2001 Census which have had identifying information removed to protect confidentiality. They are microdata files with a separate record for each individual, similar to the sort of data obtained from a sample survey. However, the sample size is much larger than most surveys thus permitting analysis of small groups and sub-national areas. The SARs allow flexible, multivariate analysis at the individual level. The SARs cover the full range of Census topics including housing, education, health, transport, employment and ethnicity. In the 2001 SAR further additional variables include religion, whether the respondent is a carer, amended ethnic group categories and more detail on qualifications. In addition, the SARs also include a range of derived variables. It is widely recognised that the SARs have provided a valuable research dataset over the last 10 years.

2.2 The legal framework of the Census

Before 1920 each individual census required an Act of Parliament. The 1920 Census Act gave statutory authority to the Registrar General for England and Wales to conduct a census every ten years. However, each census still requires an Order to be laid before Parliament giving details of the date, the topics to be covered and the population to be covered in the census. In addition, the detailed arrangements for conducting the census have to be set out in Regulations which are laid before Parliament. If topics are planned for inclusion in a census that are not specified in the 1920 Census Act then the Act has to be amended and this requires special legislation. The legislative process for the 2001 Census began with a White Paper published in March 1999 that set out proposals for the content and conduct of the census. It ended with approval of a Census Order for England and Wales on 15 March 2000, amended in December 2000 to allow the inclusion of a religion question. A question on religion was proposed in the White Paper in March 1999 subject to a change to the Census Act, which did not then provide the lawful authority for such a question to be included in a census in Great Britain.

Following devolution in 1998, parallel legislative arrangements have to be made in Scotland as responsibility for census taking now rests with the Scottish Parliament. In Northern Ireland, also, specific legislation is needed under the Census Act (Northern Ireland) 1969. Authority for the census in Wales has not been devolved although the support of the Welsh Assembly is needed. This legislative framework has a major impact on the topics included in the census and the timetable for conducting a census.

Although participation in the census in the UK is a statutory requirement, the confidentiality of the information supplied by the public is protected

by legislation. In Great Britain, the Census Act 1920, as amended by the Census (Confidentiality) Act 1991, and provisions set out in the Census Regulations lay down penalties for the unlawful disclosure of information from the census by anyone involved in taking a census. Separate legislation, the Census (Confidentiality) (Northern Ireland) Order 1991 applies in Northern Ireland.

It is unlawful for the Census Offices to pass any census information to other Government departments or any other organisation except for the purposes of the Census Act itself or the Public Records Act 1958. Under this latter legislation, the census returns are closed to public inspection for 100 years.

2.3 Why use microdata from the census

The census covers the whole population and great efforts are made to ensure as complete an enumeration as possible. The census is designed to enumerate the entire population. The size of samples of microdata drawn from the census database are not constrained by those factors which limit the size of sample surveys - most notably fieldwork costs. Samples of microdata can therefore be much larger. The SARs are samples of between 1-5 per cent of the population. The limit on sample size usually relates to confidentiality considerations rather than the cost of obtaining the sample.

Samples of census microdata are therefore considerably larger than most survey samples. This is particularly valuable for the analysis of small sub-groups of the population - for example, the very elderly aged 85 and over; minority ethnic groups; or specific family groups, such as one-parent families.

The SARs allow individual level analysis at geographical areas with a population size as small as 120,000 (the 1991 Individual SAR) and provide sufficient number of cases to allow a disaggregation of ethnic groups, for example, to distinguish Indian, Pakistani and Bangladeshi groups. By comparison with the limitation imposed by fixed, pre-defined tabulations, the SARs allow the analyst to devise new groupings for variables, new classifications (particularly for households) and to conduct multivariate analyses of individual-level data.

For further discussion of the differences between aggregate and microdata see the Collection of Historical and Contemporary Census Data and Related Materials teaching unit on Individual vs Aggregate Data at www.chcc.ac.uk/overview/faq11/frame.html.

By comparison with tabular data, microdata can support a much better specified model of unemployment and also offers some opportunity to retain the element of place. The size and unclustered sampling design of

census microdata files means that geographical areas can be identified at a more detailed level than is the case with most sample surveys. This opens up the possibility of including 'place' in the analysis through multilevel modelling methods, discussed in more detail below.

Analysis of places versus people

Where the unit of analysis is the area (at whatever spatial level is decided), then aggregate area-level data are appropriate. For example, where the aim is to identify areas with particular resource requirements, perhaps in terms of housing need or to help service providers to target resources appropriate to the needs of minority groups, then data are needed which relate to the locality.

By contrast, where the analysis is concerned with understanding individual or household-level relationships and place is of secondary concern, then individual level data are preferable. This can be demonstrated with reference to an analysis of unemployment in Britain. Unemployment shows considerable spatial variation, with very high levels often concentrated in particular areas - for example on inner city housing estates. If the requirement is to identify the areas of highest unemployment in order to target resources (e.g. Job Centres) to the most needy areas then small area data are essential. A variety of density measures for small areas have been developed and are well documented (Cole, 1993; Simpson, 1994). However, if the requirement is to identify the determinants of unemployment, making an assumption that individual-level characteristics (e.g. age, educational qualifications, ethnic group) will be of primary importance, then the application of even straightforward multi-variate analysis is hampered unless microdata are available. For example, unemployment varies with age and gender and typically shows considerable variation between ethnic groups. Educational level is known to be an important predictor of unemployment and family responsibilities may also be important. Thus a robust analysis would include at least five variables: employment status, age, sex, educational level and ethnic group, and may also require additional variables on marital status and whether or not born in the UK.

The Census Area Statistics available in Britain do not provide this level of detail in a single table. To attempt this analysis using aggregate data drawn from two or more tables will invariably raise problems of the ecological fallacy. To establish the degree of association between two variables - for example, unemployment and the level of educational qualifications - which are not available in the same table, necessitates the use of correlations based on area-level aggregate values. Seminal research by Robinson (1950) using the association between literacy rates and race in the USA demonstrated that this approach can produce seriously misleading results. In this case, there was a very high positive

correlation at the level of large administrative areas between literacy rates and the percentage of the population which was black. However, when the analysis was repeated at the individual level this association fell to 0.2. The explanation lay in the fact that Black African people tended to live in the same areas as poorly educated whites and this latter group were likely to have children with low levels of literacy. However, at the aggregate level this could not be established.

By comparison with tabular data, microdata can support a much better specified model of unemployment and also offers some opportunity to retain the element of place. The size and unclustered sampling design of census microdata files means that geographical areas can be identified at a more detailed level than is the case with most sample surveys. This opens up the possibility of including 'place' in the analysis through multilevel modelling methods, discussed in more detail below.

The ability to analyse subgroups of the population

Census microdata files may provide the only means of analysing small groups with sufficient numbers to produce reliable estimates. In Britain minority ethnic groups constitute only about 8 per cent of the total population. Because there are significant differences between ethnic groups in, for example, the level and reasons for unemployment, it is important to be able to distinguish between them. The British SARs have played an important role in identifying ethnic differentiation and highlighting incongruities between educational attainment and levels of unemployment (Blackburn et al, 1997) and the occupational level achieved (Heath and McMahon, 1997). In Britain the availability of hierarchical data has supported a nationally-representative study of family composition and partnership patterns amongst different ethnic groups (Holdsworth and Dale, 1995).

2.3.1 Level of detail

Microdata files provide detailed variable categorisations that allow the analyst to choose their own groupings or classifications. In this respect there is a marked contrast with aggregate census tabulations where all variables are precoded and include only a small number of variables in any one table. Thus the UK 1991 SARs contain 358 occupational categories on the 1 per cent file and 73 categories on the 2 per cent file. In the aggregate tables (SAS/LBS) there is only one table which provides the full 371 Occupational Unit Groups and it gives a crosstabulation by employment status and sex only. Although the Licensed 2001 SARs give less detail than the 1991 SARs they nonetheless provide much more flexibility than the published tables. The 2001 Controlled Access Microdata Samples (CAMS) provide very extensive detail. Because they are not released outside their secure setting they can include much more detail

than the anonymised files that are released. However, only non-disclosive tables or analyses are allowed out of the safe setting.

2.3.2 Choice of unit of analysis and population

Microdata files also allow the analyst to choose their unit of analysis. In Britain (as in the US and Australia) hierarchical files mean that there is a choice of working at the level of the individual, family or household. Further choices arise over the population to be analysed - for example, whether a full age range is used or restricted groups. Thus children can be selected and analysis conducted of the circumstances of the families in which they are living. Alternatively analyses can focus upon those of school leaving age or upon the elderly.

Choices also arise over analysis of those living in private households or communal establishments. The Individual SAR allows analysis of those living in residential homes, hospitals, prisons or army quarters. These choices provide the analyst with maximum flexibility - although they also require considerable care in ensuring that the most appropriate population has been selected.

In 1991 the census enumerated both 'usual' residents and also 'visitors' temporarily staying in a household. In 2001 only the usually resident were enumerated.

2.3.3 Coverage and sampling strategy

As samples of microdata are drawn from the census, they are based on a much more effective 'sampling frame' than is usually available to a social survey. Also, the fact that the census is compulsory means that, even with under-enumeration, the samples are based on very high response rates. In 1991 there was an estimated response rate of 96 per cent and these individuals formed the basis for the sample extracted for the SARs. In 2001, response was estimated at 94 per cent but the One Number Census imputed individuals and households to give a 100 per cent count of the population.

The sampling design of most social surveys includes an element of stratification and clustering - the latter in order to cut down costs. Samples of microdata from the census benefit from the fact that the census is designed to cover the entire population and therefore includes remote areas usually omitted from most sample surveys - for example, in Britain, the Highland and Islands of Scotland. In addition, the sample can be drawn using stratification based on geographical proximity, with no requirement for any clustering. This is reflected in the size of design factors - the extent to which sampling error differs from that which would be expected with simple random sampling. Therefore census microdata

files should provide a better representation of the population and also have smaller design factors than most social survey data. However, it should be noted there is some clustering at the individual level in the Household file. More details on sampling can be found in the sections below.

Research benefits

These factors make census microdata uniquely important for demographic research and population forecasting where a complete population sample is essential. The British SARs have been used in improving household projections (King and Bolsdon, 1997) in a number of ways - for example, by allowing conventional projections to be disaggregated by age, gender, marital status and household composition. They can also be used to investigate differences between types of household in their propensity to share accommodation and the impact that this might have on future household formation.

The fact that census data include the institutional population is of importance for a number of applications. Murphy and Wang (1996) use the SARs to make marital status population projections for England and Wales, where sample surveys cause problems because of the omission of the institutional population. Glaser et al's (1997) study of the health of the elderly provides a further example where the inclusion of the institutional population is essential because those with poorer health are more likely to be living in institutions, particularly if they are single. Information on migration in the census can help to understand differences in the movement of elderly people depending on their marital status and other characteristics (Al-Hamad et al, 1997). The data also allow analysis of inward migration in the last year and have been used to add important evidence to the debate over the role of social housing in restricting residential mobility and thus employment opportunities (Boyle 1995).

2.3.4 Geographical definition

The degree of geographical definition available is related to sample size and the reliability of estimates, as well as confidentiality considerations. In Britain the minimum size of a geographical area in the 1991 Individual SAR is 120,000 population, which provides much more detail than available from other microdata sources. The 2001 SAR does not provide geography below Government Office Region because of increased concerns about confidentiality. However, the 2001 Controlled Access Microdata Samples (CAMS) allow geography at local authority level.

Research value

For local government the availability of microdata at the LA level can provide considerable analytic value. For example, they provide profiles of particular client groups within the authority; they can be used to support specific policy initiatives, e.g. Care in the Community (for the elderly or those with a long-term limiting illness); and can inform transport policies - for example in providing information on modes of transport to work (Gardiner and Hill, 1997). Microdata files also provide an opportunity for alternative ways of assessing deprivation (Fieldhouse and Tye, 1996) and in advancing methods by which local authorities spending needs are assessed (Gardiner, 1996).

These factors have also made the SARs valuable to market researchers who conduct surveys in particular geographical areas and need accurate data on the socio-economic, family and household characteristics of the residents in order to produce demographic profiles and set sampling quotas.

The ability to identify geographical areas at a relatively detailed level allows 'place' to be included in the analysis by the use of multilevel modelling methods. Returning to the example of unemployment, area-level effects could be modelled most simply by assuming that overall levels of unemployment were greater for some areas than others (for given individual characteristics); but the model can be developed further by allowing the effects of the explanatory variables to vary between areas and also to include cross-level effects which summarise the relationship between an individual-level characteristic and an area-level characteristic (Jones, 1997). If the geographical areas available in the census microdata file are appropriate to the analysis, then multilevel modelling can be a very powerful method of including both individual and area-level information.

2.3.5 Addition of derived variables

The detailed individual-level information in the SARs has allowed the addition of a large number of derived variables. In the 1991 Household SAR a range of different social classifications using the detailed information on occupation and employment status were added. The same detailed information on occupation and employment status also allowed occupational status scores derived from other studies to be matched to individuals in the SARs. In 2001 detailed occupational information is only available in the Controlled Access Microdata Samples (CAMS). However, a number of derived variables have been added to these files.

In 1991 earnings information from the New Earnings Survey (NES) was added to the Household SARs. Mean hourly earnings were derived from the NES in the form of a large table broken down by variables such as age, sex, full or part-time working and region. The availability of all these variables on the SARs, coded in the same way as on the NES, allowed this 'earnings score' to be matched to all individuals in the SARs who reported an occupation. Similar information will be added to the 2001 Controlled Access Microdata Samples (CAMS) but will not be available on the Licensed Samples of Anonymised Records.

An area-based classification was added to individuals in the 1991 SARs. The 1991 Household SAR has a ward-based classification, developed by the Office for National Statistics (Wallace et al, 1995; Wallace and Denham, 1996) which assigns wards to one of 14 groups derived on the basis of their characteristics in 1991 Census data.

Another area-level classification was added to the 1991 Individual SAR this time relating to enumeration districts (ED) rather than to wards. Again, this was attached to each individual in the file. The addition of these classifications was done by ONS as it required accessing confidential information about the ward or ED in which the sample member lives. The availability of these classifications has provided an additional geographical dimension to the SARs which has supported multilevel modelling of neighbourhood effects (Fieldhouse and Tranmer (2001)). There are no area-level classifications on the 2001 SARs but the 2001 Controlled Access Microdata Samples are expected to contain the 2004 Index of Deprivation and the ONS area-level classification.

Where microdata files are organised hierarchically, as in the Household SAR where individuals within the same household are linked, there is additional scope for deriving variables which summarise the characteristics of the household or family. For example, household classifications can be derived to reflect the particular focus of the research - for example, a classification designed for a study of housing conditions might make key distinctions between single person households and couples with and without children. Identifying one-parent households might also be important. The 2001 census captured household relationships using a household matrix. The detailed information is available on the Controlled Access Microdata Sample.

2.3.6 Methodological work on coverage and quality

Census microdata files benefit from the fact that the census is a major source of population estimates and therefore a great deal of time and money is spent establishing the quality of each census. Traditionally a validation survey is conducted after each census to provide an independent check on both the quality and coverage of the census. The

Census Validation Study (Heady et al, 1994 and 1996) provided this check on the 1991 Census.

Full details of the CVS can be downloaded from www.statistics.gov.uk/about/data/methodology/specific/population/LS/resources/cvs.asp.

The National Statistics web site provides details of the coverage of the 2001 Census at www.statistics.gov.uk/census2001/methodology.asp. The Census response, the proportion of people returning a form in England and Wales, was 94 per cent. The total overall response was 98 per cent – including 4 per cent of the population estimated to be resident in households identified by enumerators but who were imputed. Through the One Number Census the final census database should hold 100 per cent of the population.

The One Number Census (ONC) aimed to integrate the 2001 Census counts with the estimated level of under-enumeration in the Census - that is the number of households and people not counted. It adjusted the Census database for the estimated undercount so that all statistics sum to 'One Number' - the national estimate of the population. The ONS claim that the results of the 2001 Census are the most accurate ever - to within +/- 0.2 per cent.

Unlike the 1991 SARs, which omitted individuals in imputed households, the 2001 SARs are drawn from the One Number Census database. Therefore some individuals and households are imputed. Flags to indicate imputed information have been included in the SARs. Therefore analysts can choose whether or not they want to include imputed information.

2.3.7 Limitations of census microdata

Whilst strong arguments have been made for the value of census microdata, it is also important to recognise that it has limitations. Firstly, it is limited in the depth of information collected. Because it is self-completion, and every household is required to fill in a form, questions must be short and simple and it cannot be unduly time-consuming to complete. This means that there is no opportunity to explore topics in the detail that is available with a sample survey. Schedules may also be completed by one member of the household on behalf of others, leading to a higher level of misreporting (Heady et al, 1996) and the self-completion nature of the schedule means that misunderstandings cannot be identified in the field. As with all data sources, it is important that the analyst recognises shortcomings in the data. The methodology checks on the census provide a good basis for assessing these.

2.4 International comparisons

Microdata is now available at an international level. The IPUMS-International database holds a steadily increasing database of census microdata from around the world at www.ipums.org/international/index.html. Data is available free of charge for research purposes only on completion of a registration form.

In the USA the Census Bureau release Public Use Microdata (PUMS) from the census. For 2000 the PUMS files contain records representing 5 per cent or 1 per cent samples of the occupied and vacant housing units and the people in the occupied units. Variables on the household file include acreage, tenure, value of housing, number of rooms, rent, utilities, number of children, income, relationship and number of vehicles. On the individual files variables include language, citizenship, place of birth, disability, earnings, education, hours worked, marital status, occupation and weight. For more information see www.census.gov/main/www/pums.html.

In Canada the national statistics office release Public Use Microdata Files (PUMFS). Three files are released: the Individual file, the Household and Housing file and the Family file. The individual file contains 345,000 records, representing 3 per cent of the families and non-family persons enumerated during the 1991 Census. The file combines details of family composition and structure. Demographic, social, cultural and economic information are provided for families, their members and for non-family persons. The individual file allows users to return to the base unit of the census, enabling them to group and manipulate the data to suit their own data and research requirements. Key variables include citizenship, class of worker, condition of dwelling, condominiums, data quality, economic family status, educational attainment, employment insurance benefits, ethnic origin, family allowances, fertility, fields of study, full time employment, hours worked, household income, household maintainers, household size, immigrant status, income, industrial classification, industries, investment income, knowledge of languages, knowledge of official languages, labour force activity, language spoken at home and religion. For more information see www12.statcan.ca/english/census01/release/index.cfm.

It is in the areas of economic policy and labour market analysis that most use of microdata in the United States and Canada has been made. In the USA, for example, economic policy-makers have made heavy use of census microdata in transport consortia and state planning departments, whilst in Canada, microdata has been used to assess the extent to which different ethnic sub-groups have managed to translate their educational qualifications into occupational advantage.

2.5 Examples of uses of the 1991 SARs

The 1991 SARs have proved an outstanding achievement for social science research. Researchers have taken advantage of the large sample sizes, the detailed geography, the wide range of socio-demographic variables and the operational flexibility. A wide range of areas have been covered including sociology, human geography, demography, economics, public health and social statistics. The availability of SARs from both the 1991 and 2001 censuses will provide a unique opportunity and challenge for studying socio- economic-cultural and demographic changes over the decade.

Research using the SARs has covered a wide range of areas in a number of different disciplines including sociology, human geography, economics and social statistics. Many users of the 2 per cent Individual SAR have exploited the large sample size and relatively detailed geography to look at social differences between sub-populations (especially ethnic groups) and between geographical areas. Users of the 1 per cent Household file have exploited the hierarchical nature of the data to look, in particular, at various aspects of household and family structure and inter-relationships within households. Users of both data sets have undertaken various types of multivariate analysis, taking advantage of the large amount of individual level information on the SARs in comparison to other census outputs.

A list of publications based on the SARs is updated annually. Some of the 'key findings' from the 1991 SARs provide potential users with information about the research use of the data. They also demonstrate that the SARs are excellent value for money!

A detailed discussion of how the 1991 SARs have been used is available at www.ccsr.ac.uk/sars/use/findings/ under the headings listed below. It is also published in Li (2004).

- Ethnic differences
- Employment and labour markets
- Household and family composition
- Migration
- Health
- Methodological developments
- Use of the SARs in Northern Ireland, Scotland and Wales
- International comparative research
- Policy use of the SARs
- Marketing and commercial use of the SARs