1991 Samples of Anonymised Records

Guide to the 1991 Samples

Applicable to:
1991 Great Britain Individual SAR
1991 Northern Ireland Individual SAR
1991 Great Britain Household SAR
1991 Northern Ireland Household SAR

This User Guide is drawn from material was previously available at http://www.ccsr.ac.uk/sars material restructured with minimal substantive edits May 2017 by Census Support, UK Data Service. Accordingly this document describes the data and data environment at the time of the first release. A small number of updates are noted in square brackets and footnotes to assist contemporary readers.

Information about the background, population bases, national estimates of the design factors of the individual SARs and the Geography of the SARs was drawn from version 2 of the User Guide to the SARs (July 1994)

Contents

Foreword to the User Guide to the SARs, July 1994

Professor Catherine Marsh, who tragically died on January 1st 1993, played a leading role in developing the scientific arguments that underpinned the decision to release the Samples of Anonymised Records. In particular, Cathie spearheaded a detailed statistical assessment of the disclosure risks which might be incurred through the release of the SARs. This work was vitally important in persuading the Census Offices that the risk was negligible. Cathie's pioneering work in assessing risks of disclosure in the SARs was recognised internationally and, in particular, by an invitation to present a paper at the conference of the International Statistical Institute in Cairo in September 1991.

In 1992 Cathie was awarded ESRC funding to establish the Census Microdata Unit at the University of Manchester, to house and disseminate the SARs and to carry out a programme of research using them. Version 1 of the SARs User Guide represented the beginning of the dissemination activities of the CMU; with the publication of this second edition, that process is well under way and we are now seeing the realisation of the research potential of the SARs, which Cathie believed in so strongly.

1. Introduction

The release of the Samples of Anonymised Records (SARs) from the 1991 Census of Population of Great Britain marked an exciting new development in the use of census data. Census users could now obtain samples of individual level data which can be exported to their own computer, instead of relying on the aggregate data contained in the Local Base Statistics (LBS), for example, or the Small Area Statistics (SAS), or having restricted access individual level data as with the OPCS Longitudinal Study [known in 2017 as the ONS Longitudinal Study - Ed]. The amount of census output supplied in tabular form is undeniably vast: for example, LBS give users access to 9 tables with approximately 20,000 cell counts, giving geographical coverage at ward level. However, tabular output is the product of attempts to conceive in advance the combinations of variables which most likely will be of benefit to census users, and almost every census user will have experienced the frustration of finding that pre-planned tables do not *quite* meet their needs. In contrast, the value of the SARs – which contain individual level responses to the 1991 Census, with unique identifiers removed in order to protect the confidentiality of respondents – lies in the fact that one does not have to be able to imagine all the uses to which they can be put: user can construct a seemingly infinite variety of cross-tabulations, using any desired variables. In essence, this allows users to manipulate census data in the same manner in which they would manipulate data from a sample survey.

Four files are available from the 1991 Census – two each for Great Britain and Northern Ireland separately. There are GB and NI Individual SARs which maximise geographical detail and have only limited information about the household in which the individual lives and GB and NI Household files which contain detailed information about all individuals in the household but have no geography below standard region.

The Individual SAR for Great Britain is a 2 per cent sample of the population. It includes 1.1 million visitors and residents in private households and communal establishments. There are over 278 SAR areas in Great Britain, based on local authorities or groups of local authorities of at least 120,000 population. The Individual SAR for Northern Ireland was released as a separate file; it also has a sampling fraction of 2% and the same

population threshold of 120,000 people. It allows ten SAR areas, based on amalgamations of co-terminus District Councils with similar characteristics. The Individual SAR files include a full range of census topics on individuals and summary information about households. Details of all variables are available in the Codebooks.

The Household SARs are 1 per cent samples of the population in Great Britain and Northern Ireland. The GB file includes 216 thousand households amounting to around half a million persons within households. The data allow linkage between household and family members. In terms of geography, the GB data includes Standard regions plus inner-London and outer-London. The Household SAR file includes a full range of census topics on individuals and derived household and family level variables. The Household SAR for Northern Ireland was released as a separate file with no geographical subdivision, it contains data on 5,255 households and 15,580 persons.

This guide provides background and methodological information for users of the 1991 Samples of Anonymised Records, including information about census collection, sampling, errors and quality and geography. The guide should be consulted alongside the codebook and glossary that is provided for each of the four files. The codebook and glossary lists the variables in each file and provides information to facilitate interpretation of the variables.

2. Background Information

2.1. The 1991 Census of Population: Great Britain

The 1991 Census of Britain was conducted on the 21st April 1991. All householders were required by law to complete the Census form with respect to each individual present or usually resident at the address on the night of the 21st April. In addition to persons in households, all persons present in communal establishments, on board a vessel or elsewhere in the area (camping or sleeping rough, for example) were also enumerated. The 1991 Census form for households contained 25 basic questions about the housing occupied by the household and about the characteristics of the individuals within that household, with additional questions on Welsh language in Wales and Gaelic language in and lowest level of accommodation in Scotland. The 1981 GB Census form had only contained 21 basic questions: the additional questions in the 1991 Census were on ethnic group, limiting long-term illness, term-time address of students and weekly hours worked. Persons in communal establishments and on board vessels were issued with an individual return form, the main difference from the form for private households being the omission of the questions on housing. The managers of communal establishments and person in charge of vessels were also required to complete a form giving details of the establishment and listing all persons present.

The 1991 GB Census returns constitute a database of around 22 million households and 57 million individuals. The responses are divided into those which are easy to code (often requiring a 'tick box' response) and those which are more difficult to code (from questions which require a written response and are therefore more costly to process). Responses to the questions in the former category are processed for the full 100% census, whilst those in the latter category are only processed for a 10% sample of household forms and a 10% sample of persons returned on forms for communal establishments. The 10% sample questions are those relating to relationship in household, hours worked, occupation, industry, workplace, journey to work, and higher qualifications. All other questions are processed at the 100% level.

There are a number of forms of output from the 1991 Census. County Reports and Monitors give information at county and district level, whilst National Reports and Topic Monitors generally give information at national and regional level (some Monitors give summary information below district level: Ward and Parish Council Monitors, for example, and Postcode Sector Monitors). Machine-readable data are released in the form of two sets of pre-specified tables: the Small Area Statistics (SAS) and the Local Base Statistics (LBS). The SAS and LBS consist of 86 and 99 tables respectively, describing the characteristics of the population, households and dwellings by individual area. The LBS tables provide greater detail in terms of the categories used in most tables; for example, variables are cross-tabulated by single years of age, rather than in five year groupings. The SAS are available for most areal units used in the census, from enumeration districts in England and Wales, output areas in Scotland and grid squares in Northern Ireland, (however, following the 1991 Census SAS for Northern Ireland will also be available at ED level and above); whilst the LBS covers areal units from ward level in England and Wales and from postcode sector level in Scotland upwards (see Cole, 1993, for further information on SAS/LBS).

The GB Census Offices also produce machine–readable datasets relating to migration and workplace, in the form of the Special Migration Statistics (SMS) and the Special Workplace Statistics (SWS). The SMS comprise ten standard tables describing the 100% set of migration flows between all origins and destinations in Great Britain and are produced as three separate sets for moves between wards, districts and customer-defined areas. The SWS comprise nine standard tables which provide information about journeys to work and the means of transport used for persons in employment by their area of residence and area of workplace, covering areas defined by customers.

1991 Census data were also been linked to the OPCS Longitudinal Study (LS)[1], a database relating to 1% of the population of England and Wales [similar studies now also exist for Scotland and Northern Ireland which also include 1991 samples - Ed] which links records from the 1971 and 1981 Censuses and which also includes information on events(births and deaths). These files have greater confidentiality risks and cannot be made available for download; instead users should contact the appropriate support unit to arrange secure access to these data. The SARs and longitudinal microdata should therefore be considered as complementary to each other, meeting two distinct needs. Anyone wanting cross-sectional 1991 Census data should use the SARs, due to the ease of access and large sample size; anyone wanting linked data, or wanting to relate information from vital events to census data (especially to calculate occupational mortality) should use the Longitudinal data.

In 1991, the Economic and Social Research Council and Information System Committee of the Universities Funding Council (now known as JISC: the Joint Information Systems Committee) jointed together a major census programme:

- To finance the provision of census data as a research resource to the academic community;
- To provide support for access to the datasets;
- To provide training in the use of census data;
- To promote census-based research

[1] The OPCS Longitudinal Study is, at the time of publication of this version of the user guide called the ONS Longitudinal Study and has been updated to additionally link the 2001 and 2011 Censuses.

- And, to provide support for the development of further research products arising from census output

This programme, coordinated by Professor Philip Rees at the University of Leeds, contributed to the funding of the Longitudinal Study Support Programme at City University and the ESRC Data Archive at the University of Essex. It also funded the establishment of the Census Dissemination Unit (supporting LBS, SAS, SMS and SWS) and the Census Microdata Unit, both at the University of Manchester with Manchester Computing Centre playing a key role.

2.2. The 1991 Census of Population: Northern Ireland

The 1991 Census of Population for Northern Ireland was conducted on the same date as the GB Census (21 April). The Northern Ireland Census form contained 27 questions and in the interest of maintaining compatibility of statistical information with the UK the questions were modelled closely on those contained in the GB Census forms. The ethnicity question, introduced in the GB for the first time in 1991, was excluded from the Norther Ireland Census, but in keeping with long established practice the latter included a voluntary question on religion. The Northern Ireland Census included some additional questions: a question on the number of children born alive in marriage, which was last asked in 1961, was reinstated; under the 'economic activity' question a category was included to cover 'unpaid work in a family business, including a shop or farm'; and the amenities question was widened to include water supply and domestic sewage disposal. Also, and for the first time in a Northern Ireland Census, a question was included on knowledge of the Irish language. The wording of this question was identical to the wording of the Gaelic question contained in the 1991 Census for Scotland.

These differences not only reflect the different needs of administrators and policy makers in the province, but also highlight the historical context of the census in Northern Ireland. Up until 1972, the Northern Ireland Census was authorised by Stormont rather than Westminster, and Stormont had complete discretion in matters of the timing, logistics and content of the census, although in practice there was a large degree of cooperation between the two Census Offices for Great Britain and the Census Office for Northern Ireland. The 1971, 1981 and 1991 Censuses were each authorised by the Census Act (Northern Ireland) 1969, which makes provision for a census in the province from time to time by Order of Council, as long as a period of at least five years has elapsed between censuses.

The 1991 Northern Ireland Census returns constitute a database of approximately 530,000 households and 1.5 million individuals. Standard census output for Northern Ireland is generally released at a level of administrative and/or electoral areas in the form of a Summary Repot and topic reports. The subjects covered in the Summary Report include area, population, households, age, sex, marital status, birthplace, religion , dwelling type, tenure, household size, occupational density, amenities, cars, economically active persons, long term illness, term-time address of students and schoolchildren, academic qualifications and Irish language. Greater detail on these subjects is contained in the specific topic reports. Since 1971, to meet user requirements it has also been possible to provide small area statistics by grid squares (100 metre for urban areas and 1 kilometre for rural areas).

2.3. Background to the Release of the SARs

British academics have made attempts to obtain a sample of individual-level records from the Census on more than one occasion prior to the successful release from the 1991 Census. Previous requests for SARs drawn from GB census data, however, floundered on whether the provision of data in such a form would constitute a 'statistical abstract', and thereby satisfy the requirements of the 1920 Census Act. Precedents for the release of 'public use' data have existed for some time: in the USA, for example, samples of microdata were released retrospectively from the 1960 census and have been routinely produced since, whilst Canada and Australia followed in the 1970s and 1980s respectively. Microdata from the US Census have been made available as a 5 percent count/county equivalent file, a 1per cent metropolitan area file and a 3% sample of the elderly. Microdata from the most recent Canadian Census will be released as an individual file, a household and housing file and a family file, whilst microdata from the latest Australian Census has recently been released in two samples, one containing finer geographical detail at the expense of finer categorical detail in other areas. Census Offices in the majority of European countries have not as yet released full samples of anonymised records, although some (the French Census Office, for example) have released a limited set of variables and cases, and others (such as the Italian Census Office) have made microdata available to regional offices. Spain and Luxembourg have released microdata for the exclusive use of government departments, whilst academics in both Norway and Sweden can obtain strictly limited access to anonymised census microdata (Middleton, 1993).

In 1987, at the request of the Census Office, a working party of the Economic and Social Research Council conducted a survey amongst academic users of census data on a number of census-related issues, including the desirability or otherwise of having individual level data made available from the 1991 Census. Over a third of the 220 respondents said that access to such data would make a major improvement to their research, whilst a further third felt that it would be highly desirable for, or of interest to , their research (Marsh et al. 1988). A formal request was subsequently submitted to the OPCS concentrating on the benefits of releasing such records, the uses to which they would be put, and giving an assessment of the confidentiality risks involved in their release.

That there was a shift in the official attitude towards the release of SARs became evident on the publication of the 1991 Census White Paper in 1988. The White Paper firstly noted that the legal advice now held that SARs could be deemed statistical abstracts, and then went on to say that 'requests for abstracts in the form of samples of anonymised people and households would also be considered, subject to the overriding need to ensure the confidentiality of individual data' (OPCS & GRO(S), 1988).

The Registrars General for England, Wales and Scotland announced in July 1990 that they had agreed in principle to the release of SARs and detailed work then began on developing the statistical specification. Following receipt of an independent report on the confidentiality aspects by Professor Tim Holt of the University of Southampton, it was announced in March 1992 that two SARs from the censuses in England and Wales and in Scotland would be produced for purchase by the ESRC: a 2% sample of individuals in households and communal establishments and a 1% sample of households and the individuals contained in those households. As a result of competitive application, the University of Manchester was awarded an ESRC contract to house and disseminate the SARs. The Census Microdata Unit (CMU) was duly set up in March 1992 under the directorship of the late Professor Catherine Marsh, who had played a leading role in

securing the release of the SARs. The data were finally released to the CMU in August 1993 and became available to users in October 1993. It was known at the time that there were problems with two variables (distance to work and distance moved), and a new corrected version of the SARs was provided by OPCS in January 1994. Further minor errors in the data were corrected later in 1994.

A similar request was made with respect to the release of SARs from the Northern Ireland census, which were released to the CMU in May 1994 and made available to users in July 1994.

2.4. Disclosure Control Measures in the 1991 SARs

There is a legal requirement for everyone to complete a census schedule. However, the Census Offices also have an obligation, under the 1920 Census Act, not to disclose any identifiable information that has been provided in the census. It is therefore of the paramount importance that samples of anonymised records from the census do not pose any threat to this confidentiality undertaking.

In order to address the risk of disclosure, the Economic and Social Research Council Working Party set out a four-stage process through which disclosure could occur. The process was premised on the assumption that, before disclosure could occur, an individual or household would have to be correctly identified. It was assumed that the most likely way for this to occur was by matching variables in the SAR with the same information on another, external file which held the identification of the individual or household.

The four steps necessary for disclosure were:

(1) Key variables in the microdata sample would have to be recorded in a compatible way on an outside file. If key variables are not recorded in the same way, or contain errors, then correct matches are unlikely.

(2) The individual in an outside file would have to be selected in the microdata sample before a match is possible.

(3) The individual's combination of values on the key variables must be unique in the population - otherwise an apparent match with a member of the microdata sample could, in fact, be with a 'statistical twin'.

(4) The person attempting to make the match would need to be able to verify uniqueness in the population - for example by having a list of the entire population on the key variables and thereby being sure that the match is, indeed, correct.

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 was negligible. Details of the calculations made at each stage are given in Marsh, Skinner et al. (1991). The arguments put forward were important in persuading the Census Offices to release the SARs, suitably modified to protect anonymity where this was felt to be at risk.

An independent technical assessor, Professor Tim Holt, was also appointed to advise the Registrars Generals on the confidentiality aspects of the release of SARs. The conclusions of his report were cited in a written Parliamentary Answer on the 11th March 1992 in reply to a question asking whether the SARs were to be released or not:

"As a reasonable statistical judgement, the risk of disclosure is negligible for the large

majority of the population and thought to be extremely small for the remainder. International experience indicates that the levels of risk are consistent with a decision to go ahead with the release of the SARs, and I so advise."

In the next section the various disclosure protection measures taken are described.

2.4.1. *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 also reduces the chances that anyone would make an attempt at identification by this means. The two SARs (totalling 3 per cent together) 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.

2.4.2. *Restricting geographical information*

One of the key considerations which may affect the risk of identifying an individual or household is the geographical level at which data is released.

Empirical work and comparisons with SARs released in other countries showed that a sensible level for release would be areas with a population size of at least 120,000 in the individual (2 per cent) SAR. This level of geography allowed the majority of British local authorities to be separately identified. Smaller local authority districts (under 120,000 population) were grouped to form areas over 120,000. Only one geographical scheme was permitted in order to avoid overlaps where the difference between two areas could lead to the identification of sub-threshold area. The geography of the Northern Ireland 2 per cent Individual SAR is based on combinations of District Councils into ten geographical areas, each area again having a minimum resident population of 120,000. Areas have been defined in order to combine district councils with similar characteristics.

The one per cent household SARs, because of their hierarchical nature (i.e. records for the household and all its members), are more of a disclosure risk. For this reason it was decided that, for this SAR, the lowest geographical level 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. The smallest region, East Anglia, has a population of about 2 million. The 1 per cent Household file for Northern Ireland has no geographical subdivision.

Other geographical information obtained from the Census – usual address of visitors to the household, workplace address, students' term-time address and address one year ago is heavily restricted, limited to standard regions or to a same SAR area / different SAR area identifier.

Part 5 of this guide details on the geography of the GB SARs, including a map of the 278 SAR areas and details of the geography of the Northern Ireland SARs, with a map of the ten SAR areas for Northern Ireland.

The geographical ordering of records in both SARs do not reflect their geographical ordering within the ONS database. Although sampled with households grouped by county and enumeration district in England and Wales and by region and output area in Scotland, once selected, the records have been scrambled. This prevents any possibility

of tracing individuals or households back through a region or district.

2.4.3. *Suppression of data and grouping of categories*

Some alterations were made to the data to reduce the number of rare and possibly unique cases. Information which is unique in itself, such as names and addresses, is, of course, omitted altogether (neither is it included on the ONS Census database in the first place), whilst the precise date of birth of individuals has been suppressed (age is based on the number of completed years of age at the time of the Census).

With some variables, categories with small numbers have been grouped, either across the entire range of the variable or at the extremes, particularly at the upper extreme (for example, for those aged over 90) where grouping high values is known as ‘top coding’.

The rule used to decide the level of detail to be released was that, on average, the expected sample count would be at least one for each category of each variable at the lowest geographical area permitted on each SAR. This rule was applied to each census variable.

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 topcoded above 14 and the number of persons in the household was topcoded above 12. With age, 91 and 92 were grouped, 93 and 94 were grouped and 95 and over was topcoded.

Generally, less detail was released on the two per cent individual SAR, because of the lower level of geography. For example, occupation was reduced to 73 categories whereas in the household file it was coded to 358 categories. The other area of special concern was geographical information on workplace and address one year before census; this was heavily grouped before release.

Some additional restrictions were applied to certain occupational groups which were considered a particular risk because of being in the public eye - for example actors, professional sportsmen and women and politicians.

Large households were also seen as a disclosure risk in the household sample. Therefore for households containing 12 or more persons no information about the individuals in the household is given. In fact, only 28 households in the sample contained 12 or more persons

2.4.4. *User obligations*

Finally, SAR users have to give an undertaking not to obtain or derive information relating specifically to an identified individual or household, nor claim to have derived such information. Due to the uniqueness of the SARs in British census history, it is extremely important that these conditions are met. Any breaches of the undertaking will result in the recall of the data.

2.5. Sampling in the 1991 SARs

As previously noted, the coding of the 1991 Census for Great Britain was divided into two stages. Easy to code information, such as sex, date of birth, marital status and country of

birth, was processed for all forms and then a 10 per cent sample 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 as the base from which to draw the SARs.

The sampling for the SARs was divided into two stages, with the one per cent Household file selected first. All fully-coded household forms were ordered geographically with the lowest level the enumeration area (about 200 households). Households were then grouped into batches of 10 and one household selected at random from each batch. All sampled records were then scrambled before release to prevent households being traced by their geographical ordering.

The two per cent Individual sample was then drawn 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 the same household. 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 within each SAR area before being released.

In Northern Ireland all records were 100 per cent coded. The NI 1 per cent Household sample was selected first by stratifying households within enumeration districts and District Councils into groups of 100 households and then selecting one household at random from each group. The hierarchical household SAR contains 20,833 records in total: 5,255 households and 15,578 persons within those households. The variables are broadly similar to those within the GB household file (bearing in mind the differences which exist between the two Censuses) whilst, as in the GB file, there are no individual records released for households containing 12 or more individuals. The Northern Ireland census database does not hold information as to which family an individual belongs, so unlike the three-tiered structure of the GB Household SAR, analysis is only possible at the level of the household and the individual.

The NI 2 per cent Individual sample was selected by stratifying the remaining individuals into groups of 99 and by choosing two individuals at random from each group. Individuals in communal establishments were stratified geographically into groups of 50 people and one person was chosen at random from each group. There are 31,967 individual records on 10 area files, with information provided on an individual's resident status (present resident, absent resident or visitor). As with the Household file, the variables contained within the Northern Ireland 2 per cent Individual file are broadly similar to those in the GB 2 per cent file.

As with the GB SARs, to prevent any possible geographical tracing within a SAR area, the files were scrambled before release.

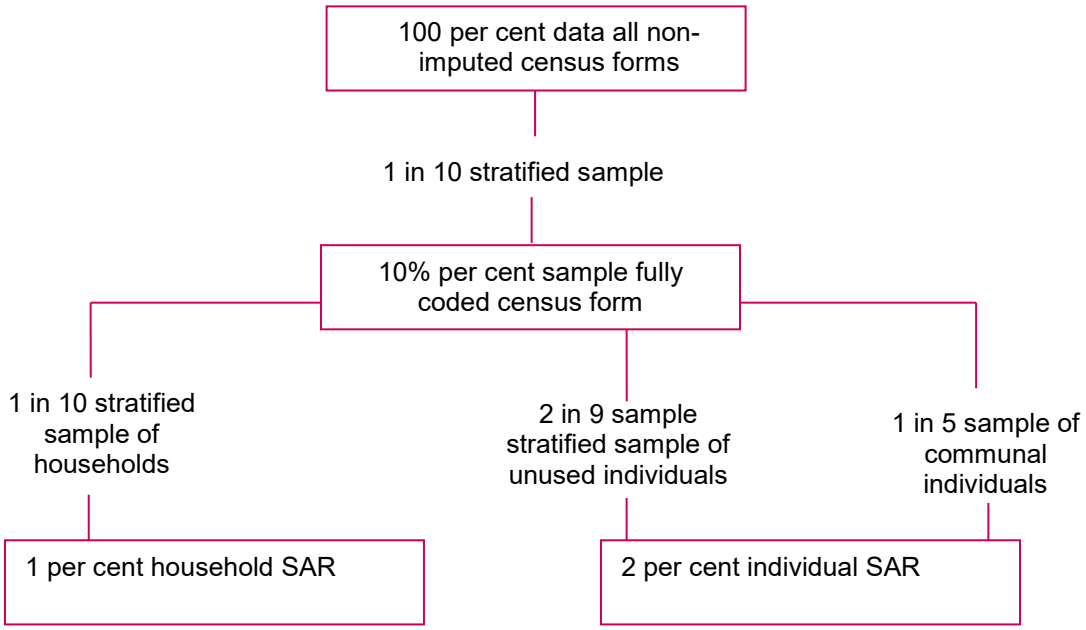


Fig 1: Sampling in the 1991 SAR

Variables in the SAR files show the effects of both stratification and clustering. Attributes that tend to be common across areas will be affected by stratification (for example, local authority housing tenure) and will therefore have a lower sampling error than that for a simple random sample. Other variables, where values tend to be the same for all household members, will be affected by clustering, leading to larger than expected sampling errors. For example individuals within the same household are likely to have the same ethnic group and social class. The effect of clustering is more pronounced for individual level variables in the household file, as all individuals in each household are selected for this sample.

Staff at the Census Microdata Unit carried out a programme of validation on the two SAR files. This took three forms: firstly, a series of logical checks was carried out to ensure that information within each individual record and household set of records was logically consistent (errors such as married under 16 year olds would have been picked up as part of census office edit checks; CMU checks were therefore limited to SARs-specific issues such as checking that households with more than 12 members did not include person records for example). Secondly the data were (i) compared with the population characteristics (ii) providing a measure of sampling variability and (iii) measuring differences between SAR and published outputs.

2.6. Differences across UK Countries

The 1991 Census form was very similar in England, Wales and Scotland. In Northern Ireland there were more differences in questions – e.g. religion was asked and fertility of married women, and differences in coding procedures – for example data were 100 per cent coded. As for GB, two non-overlapping samples are available. As far as possible, similar derived variables have been added to the Northern Ireland SARs.

2.6.5. *Differences in the treatment of family variables.*

While the majority of questions and concepts are the same there were some differences between Northern Ireland and Great Britain. For example in GB a cohabiting couple living on their own would be treated as a family, while in Northern Ireland the couple would need to be married to constitute a family. Accordingly, an unmarried cohabiting couple with children would have been classified as a lone parent family with others. Similarly, 'cohabitant of son/daughter' and 'boarder/lodger' are not distinguished in the Northern Ireland relationship to head of household question, while they are in the GB.

There are additional differences in the treatment of families between GB and NI with respect to processing family data. In GB the allocation of individuals to families was done using a complex computer algorithm, whereas in Northern Ireland the number of families is determined at the coding stage. The latter allocated individuals to a much smaller range of family and household types.

2.6.6. *Distance to work and previous address.*

The GB SARs make use of the Central Postcode Directory (CPD) to calculate the distance of move of migrants and distance to work. However, Northern Ireland rely on the grid square references. Additionally grid square references are only available for the place of employment where there are more than 25 employees. These variables are not available in the NI SARs.

2.6.7. *Edit and Imputation*

The remaining processing difference relates to editing and imputation procedures. Computerised editing was used for the GB Census to identify inconsistencies and missing values and then to impute valid, consistent answers (Mills and Teague 1991). In Northern Ireland only clerical processing was used. Some inconsistencies were resolved manually at the coding stage and clerical procedures give rules for handling missing answers. The approaches will, however, result in different values being imputed where data is missing: for example, if there is a missing valid for the number of cars in a household, the computerised system will produce an imputed value based on the number of people in the household, the tenure of the household, and whether the accommodation in a permanent or non-permanent building, whereas the clerical procedures used in Northern Ireland would result in a value of zero being imputed.

3. Accuracy of estimates from the 1991 SARs

3.1. Introduction and overview

Estimates prepared from the Samples of Anonymised Records are based on a sample of the 1991 Census data. They are estimates of the actual figures that would have been obtained from a complete enumeration of all residents. These estimates are expected to be different from complete figures because they are subject both to sampling errors and non-sampling errors. They will not necessarily be the same as those published in census reports. This section of the user guide discusses sampling and non-sampling errors in some detail and suggests how the user should assess these errors in practice. The advice can be summarised as follows:

- users should calculate confidence intervals to reflect the sampling error attached to estimates calculated from the SARs
- users should be aware of the most likely deficiencies in the quality of census responses and include relevant cautions in reports based on SARs data
- users should adjust figures from the SARs to take into account the incomplete coverage of the population, particularly where totals of population categories have been estimates rather than ratios or percentages.

3.1.1. Sampling error

Because the SARs are based on a random sample, estimates based on them may differ somewhat from the figures that would be obtained from processing all the census records; they may also differ from the estimate that would have been obtained from processing a different sample of the same size drawn in the same way from the census records.

For a limited number of variables it has been possible to compare the estimate from the SARs and the value that is given by all census records that the SARs are drawn from. For Great Britain these comparisons are given in the table below. Throughout the table, the population base includes present and absent residents but excludes visitors, imputed absent households and residents in imputed absent households. As one might expect with such large samples, the SARs closely represent the population from which they were drawn. Conversely, the smaller the size of the sample the greater the tendency of estimates to differ from corresponding values for the entire population. Consequently, estimates derived from the SARs for sub-groups of the population or single SAR areas will tend to deviate more from the 100 per cent statistics.

The deviation of a sample estimate from the census value is called the sampling error. The standard error of a sample estimate is a measure of the variation (the standard deviation) of the sampling error across all the possible samples and thus is a measure of the precision with which an estimate from a particular sample approximates the census value.

Table 1: Characteristics of the SARs and the population from which they were drawn Great Britain, percent.

Individual Characteristics

	% OF ALL RESIDENTS		COMMUNAL ESTABLISHMENTS	
	Individual SAR	Census population	Individual SAR	Census population
Male	48.4	48.4	41.7	41.2
Female	51.6	51.6	58.3	58.8
Age 0-15	20.2	20.2	3.4	3.7
16-17	2.5	2.5	1.2	1.3
18-29	18.1	18.1	21.6	21.3
30-44	21.3	21.2	9.9	10.1
45 up to pensionable age	19.3	19.3	8.6	8.9
Pensionable age	18.7	18.7	55.2	54.7
Single	41.0	41.1	47.3	48.0
Married	46.9	46.8	12.5	12.7
Widowed/Divorced	12.1	12.1	40.1	39.3
With llt illness	13.1	13.1	63.5	63.3
In employment	44.1	44.3	21.6	22.1
Unemployed	4.6	4.5	3.7	3.6
Economically inactive	31.2	31.1	71.3	70.6
White	94.6	94.6	94.3	94.5
Other ethnic groups	5.4	5.4	5.7	5.5

Household Characteristics

	% OF RESIDENTS IN HOUSEHOLDS		% OF HOUSEHOLDS	
	Individual SAR	Census population	Household SAR	Census population
One person in household	10.6	10.6	26.3	26.3
Owner occupied	69.9	70.0	66.4	66.7
Rented privately (exc with job)	5.5	5.5	7.2	6.9
Rented from a housing association	2.4	2.4	3.2	3.1
Rented from a local authority, new town or Scottish Homes	20.0	20.0	21.3	21.4
Lacking or sharing use of a bath/shower and/or inside WC	0.74	0.75	1.3	1.2
No central heating	16.8	16.8	18.8	18.8
No car	24.9	24.9	33.3	33.1
Lone parent	n/a	4.15	3.7	3.7

Sources. Individual SAR, Household SAR, LBS (Tables 18 and 19 for imputed households, deducted from equivalent cells for 100 per cent data in other LBS tables). The base in each case excludes imputed households and residents in them. Crown Copyright.

The sample estimate and its estimated standard error permit the construction of interval estimates with prescribed confidence that the interval includes the true population value.

3.1.2. Non-Sampling error

In addition to the variability which arises from the sampling procedures, both sample data and the full census data are subject to non-sampling error. Non-sampling error may be introduced during any of the complex operations used to collect and process census data.

Non-sampling error may affect the data in two ways. Errors that are introduced randomly will increase the variability of the data, and should, therefore, be reflected in the standard error discussed below. Errors that tend to be consistent in one direction will make both sample and 100 per cent data biased in that direction. For example, if respondents consistently tend to under-report the number of cars available to their household then the resulting counts of households by number of cars will tend to be understated for the multi- car households and overstated for the no-car households. Such biases are not reflected in the standard error.

Sources of non-sampling error include:

a) Quality of response

Respondents to the census may misinterpret census questions or for other reasons complete the census form incorrectly. The census form requests that the head or joint heads of the household, or other adult over 16, completes the form on behalf of all members of the household. The Census Validation Survey (CVS) carried out by OPCS shortly after the 1991 Census assessed the quality of responses to the census.

b) Incomplete coverage of the census

Every census misses some people who are particularly difficult to enumerate, in spite of the thorough census field procedures designed to enumerate the entire population (evaluated in Clark 1992). The age-structure of those missed by the census has also been estimated and is significantly different from the age-structure of the population as a whole.

c) Transcription and coding errors, missing data items

During the processing of census forms, transcription and coding errors can occur. Missing items for persons on a completed census form are imputed (estimated) by the Census Offices. Corrections are made to some inconsistent data, such as persons reported married but aged under 16. Mills and Teague (1991) provide a description of the processing of census forms and the imputation of these types of missing or inconsistent data.

d) Data Modification to ensure confidentiality

In the 100 per cent tabular output of Local Base Statistics and Small Area Statistics for areas within local authorities, an additional source of error was purposefully introduced by Census Offices to provide additional protection against the identification of individuals (Census User Guide 48; Cole 1993). Counts in some cells of the tabulations are slightly adjusted; the cells that are adjusted are not known to the user. *However, no such adjustment is made to the sample data in the 10 per cent tabular output or in the SARs*. Other methods, described in Section 1.4, reduce the already negligible risk that individuals can be identified from records in the SARs.

3.2. Standard Errors and Confidence Intervals

The complex sampling design described above has implications for the estimation of sampling errors on both the individual and household file. The 1 per cent household SAR approximates to a simple stratified random sample of households, although counts of individuals in the household file are subject to the effects of clustering. In the 2 per cent individual file there are two potential sources of clustering which arise in the sampling process. First individuals are clustered into households in the selection of the 10 per cent sample and second, the removal of the household SAR from the 10 per cent sample implies a further clustering into households (Dale and Marsh, 1993). Nonetheless, preliminary work suggests that the 2 per cent SAR approximates to a simple random sample.

The method described here for estimating standard errors of estimates from the SARs involves two simple stages. The first stage calculates the unadjusted standard error, using formulae that apply to simple random samples. The second stage multiplies the unadjusted standard error by an appropriate design factor. This is the factor by which sampling errors must be multiplied in order to compensate for the effect of clustering or stratification in the sampling process. The design factor approximates the ratio of the standard error from the actual sample design to the standard error from a simple random sample. In practice the steps are:

1. Calculate the unadjusted standard error from the appropriate formula.
2. Multiply the unadjusted standard error from step 1 by the design factor appropriate to the characteristic (e.g. unemployment status, or age).

The design factor that should be applied may be more or less than 1.0. If there is stratification in the sampling process the sample should be more representative than a simple random sample and the design factor will be less than one. Clustering will cause sampling errors to be larger than those found with simple random sampling and the design factor will be greater than one.

Preliminary estimations of design factors have been made using two different methods, the first using sampling point information (for the household file); the second comparing differences between expected and observed errors (for the individual file).

3.2.3. Design factors for household characteristics : 1 per cent household SAR

Assumption of a design factor of 1.0 (i.e. using the unadjusted standard errors as if the sample was a simple random one) is unlikely to be far wrong when using household characteristics from the 1 per cent SAR. At worst, a slight over-estimate of the sampling error may result, as household level variables are subject to stratification effects and estimated design factors (including those relating to particular members of the household, for example the social class of the head of household) are slightly less than unity.

3.2.4. Design factors for individual characteristics : 1 per cent household SAR

For analyses of individual characteristics from the 1 per cent household SAR, assuming simple random sampling may be misleading because clustering effects mean that sampling errors may be seriously under-estimated. This is because this SAR includes all individuals in each sampled household and for variables such as ethnic group, country of birth, migrants, qualifications and social class, there is a tendency for individuals in the same household to have similar characteristics. The effect of household clustering could probably be ignored however for estimates of subgroups of which there is usually no more than one person per household, such as women aged over 80.

The largest effects are for ethnic group. Preliminary estimates are as follows:

Ethnic Group	
White	1.84
Black Caribbean	1.60
Black African	1.83
Black Other	1.51
Indian	1.99
Pakistani	2.27
Bangladeshi	2.37
Chinese	1.87
Other Asian	1.83
Other Other	1.60

3.2.5. Design Factors in the 2 per cent Individual SAR

Design factors estimated for the individual SAR are based on a comparison of the difference between the SARs and 100 per cent Census data across the 278 SAR areas (having subtracted residents in wholly imputed households) and the sampling errors which would be expected from simple random sampling. The method is described in more detail in CMU Occasional Paper 2.

Individual and household level variables on the individual file are less likely to be subject to clustering and may benefit from stratification. Most design factors deviated very little from unity, many being less than one. Again the largest design factors are for ethnic group, though the effects are much smaller than on the household file.

Ethnic Origin

	LA level calculations	National calculations
White	1.15	1.01
Black Caribbean	1.00	1.05
Black African	1.06	1.08
Black Other	1.04	1.07
Indian	1.26	1.05
Pakistani	1.20	1.11
Bangladeshi	1.04	1.18
Chinese	1.19	1.11
Other Asian	1.02	1.10
Other Other	1.30	1.07

The second (national) calculation will normally give more accurate design factors as it based on a method of calculation that it is based on a method similar to that used to produce design factors for the Household SAR. It is based on a comparison of individuals grouped into pairs of consecutive households within counties. Individuals omitted within communal establishments are grouped into consecutive pairs within counties.

This method allows for clustering of individuals in the same household and also for most of the stratification present in the sample design. Since there are substantially more "degrees of freedom" involved in this method of estimating one may expect this to be a more precise method of calculation.

Having calculated the standard error for a SAR estimate, it will often be appropriate to go on to calculate a confidence interval for the estimate. These are discussed below. Users should also read the notes below which give further advice on the use of standard errors. Worked examples are given throughout this discussion of standard errors and their use.

Generally, use of the 2 per cent individual SAR will minimise sampling errors for individual level analyses whilst the household file is the most appropriate for the analysis of household characteristics.

3.2.6. Calculation of standard errors

The means of calculating the unadjusted standard errors for four common statistics are given here. The derivations can be found in many statistics textbooks and most statistical software will calculate them as part of their standard output.

Statistic	Value	Approximate standard error
Sample cell count	c	SE(c)=sqrt(c(N-c)/N)
Scaled cell count	C=f*c	SE(c)=f*SE(c)
Sample cell proportion	Pr=c/n	SE(Pr)=sqrt(Pr(1-Pr)/n)
Sample cell percentage	Pe=100*c/n	SE(Pe)=sqrt(Pe(100-Pe)/n)

Examples of the statistics, and definitions:

c the number of non-white textile workers in Yorkshire and Humberside region.

C=f*c that number scaled to the total census enumerated population. In this case f is 50 or 100 for the individual or household SAR respectively.

N the total number of records in the SAR in the Yorkshire and Humberside region, irrespective of industry or ethnic group. In general, N is the total number of records in the SAR for the area concerned; where a characteristic of the population in communal establishments is being counted in the individual SAR, N is the total number of records from communal establishments in the area concerned.
Where N is very large compared to c (N more than 30 times c), the formula for SE(c) can be replaced by the approximation SE(c)=sqrt(c) and SE(c)=f*sqrt(c).

Pr The number of non-white textile workers in the region (c) as a proportion of all non-whites in employment in the region (n).

Pe The number of non-white textile workers in the region (c) as a percentage of all non-whites in employment in the region (n).

The standard error of the SAR statistic is then derived by multiplying the unadjusted standard error from these formulae by the appropriate design factor.

Examples of calculation of standard errors

(a) The percentage of the population of Newham who are of Indian ethnic origin. The percentage of the sample who are Indian in Newham is 13.8 per cent.

The unadjusted standard error is

Unadjusted SE(Pe) = sqrt(13.80*(100-13.8)/4,179) = 0.53

The estimated design factor for Indians on the individual file is 1.26. The standard error for this SAR percentage is therefore

Standard error (Pe) = 0.53*1.26 = 0.67

(b) The number of renting households in Britain with a person under pensionable age having a limiting long-term illness, from the household SAR, scaled to a total for all households enumerated in the census.

If the total number of such households in the household SAR is 289, it is scaled by 100 (the household SAR sampling fraction) to estimate a total in Britain of 28,900 such households. There are 215,789 household records in the household SAR in all, so the unadjusted standard error of the estimate of 28,900 is

Unadjusted SE(c) = 100*sqrt(289*(215,789-289)/215,789) = 1,699

Note that the number c=289 is very small compared to the overall number of records N=215,789, so a very similar result would be achieved using the approximation referred to on the previous page,

Unadjusted SE(c) = 100*sqrt(289) = 1,700

From the discussion in the previous section, the design factor for household characteristics from the household SAR may be taken to be 1.0, so in this case the standard error requires no further adjustment.

3.2.7. Confidence intervals and inferences based on the SARs

A sample estimate and its estimated standard error may be used to construct confidence intervals around the estimate. These intervals are ranges that will contain the true population value of the estimated characteristic, with a known probability.

For example:

1. With approximately 68 per cent probability, the interval from one standard error below the estimate to one standard error above the estimate contains the true value.
2. With approximately 90 per cent probability, the interval from 1.6 standard errors below the estimate to 1.6 standards error above the estimate contains the true value.
3. With approximately 95 per cent probability, the interval from two standard errors below the estimate to two standard errors above the estimate contains the true value.

The intervals are referred to as 68 per cent, 90 per cent, and 95 per cent confidence intervals, respectively.

Example
Using the earlier example, the standard error of the 28,900 households in Britain with someone below pensionable age with a limiting long-term illness was estimated to be 1,698. Thus a 95 per cent confidence interval for this estimated total is estimated as:

(28,900 - 2*1,698) to (28,900 + 2*1,698), or 25,504 to 32,296

3.2.8. Other notes on standard errors

A standard sampling theory text or the explanatory guide to the user's statistical software should be helpful if the user needs more information about confidence intervals and non-sampling errors. These should be consulted for details of standard errors for sums, differences and ratios of estimates from the SARs.

Zero estimates
When the proportion, percentage, or cell count is zero, the formulae in section (ii) above give estimated standard errors of zero. While the magnitude of the error is difficult to quantify, estimated percentages and totals of zero are still subject to error.

The effect of non-sampling error on the standard errors and inference using confidence intervals.
The estimated standard errors given above do not include the variation due to non-sampling error that may be present in the data. The standard errors reflect the effect of simple response variability, but not the effect of systematic errors introduced by enumerators, coders, or other field processes. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence in estimating the true population value of a characteristic. One of the most important sources of error that might additionally affect the accuracy of confidence intervals is bias arising from missing records, discussed below.

3.3. Quality of Census Responses

One characteristic of the SARs is that the accuracy of responses contained in them is determined almost wholly by the accuracy of the responses given by residents themselves. There has been no data modification or perturbation and imputed records for wholly absent households have not been included in the SARs; only data that was missing from a returned form has been imputed, by the 'hot deck' procedures described in Mills and Teague (1991), for the 100 per cent coded variables.

A check on the quality of responses in the Census was one of the aims of the Census Validation Survey (CVS) carried out very soon after the 1991 Census. The CVS, which seeks to establish the quality of both responses to, and coverage of, the Census, is based on a sample of around 6000 households in over 1200 enumeration districts, and was administered by means of individual interviews held between six weeks and three months after Census day (Wiggins, 1993). Results from the 1991 Census Validation Survey: quality report (Heady et al, 1996) HMSO contains full details of the quality of responses.

The vast majority of questions in both Censuses yielded a gross error rate of less than

5%. Across both Censuses, however there were three questions which generated a higher error rate: number of rooms in the household (28.3%), economic position (10.9% in 1991) and means of travel to work (8.1%), all of which had very similar gross error rates in 1981. In addition, two derived variables yielded relatively high error rates in 1981: social class (13%) and socio-economic group (16%). These rates mainly reflect errors made by form fillers, but also take account of errors made in the processing of the results which were not subsequently corrected by the editing process. The first three variables will be briefly considered, whilst the first results of the quality of responses to the new question on ethnic group will be summarised.

Number of rooms

Inaccurate recording of the number of rooms in a household arose in 1981 largely from confusion about whether or not to include the kitchen as a room and resulted in an *under*count by 1 room for households with a small number of rooms and an *over*count by 1 room for households with a larger number of rooms. Instructions on the 1991 form gave greater detail to form fillers on which rooms should be included in the count, but resulted in only a slight reduction in the gross error rate.

Economic position

Inaccurate responses in 1991 to the question relating to economic position were higher for women than men (13.6% compared with 7.8%). In 1981 the error rate arose in the main from difficulties in distinguishing part-time from full-time employment, and distinguishing between the categories included under the general heading of 'economically inactive'. The 1981 Post Enumeration Survey (PES) revealed for instance that there had been some confusion between the categories 'housewife' and 'retired' (with 'housewife' always given priority over 'retired' in the coding process). The 1991 Census form replaced the 'housewife' category with a 'looking after home and family' category, yet it has clearly not had the desired effect in clarifying the question.

Given that the SARs allow for detailed analysis of an individual's *secondary* economic position, it is worth noting the comments of the 1981 PES on multi-ticking of the economic position question. Multi-ticking was done in the main by 'housewives' (the category used in the 1981 Census) and by people in part-time employment. However, it was expected – and indeed confirmed by the PES – that all the possible circumstances relating to an individual's economic position would be understated. There was evidence that many 'housewives' who worked part-time did not tick 'part-time employment' as a secondary category because they did not regard their jobs as 'proper' part-time jobs or did not consider them worth mentioning because of the few hours and/or low financial returns involved (mail order agent or babysitter for example). The PES found that 7% of cases in the 'other inactive' category (mainly made up of 'housewives') in fact had a part-time job at the time of the census.

Means of travel to work

In 1981 problems in this question arose as a result of ticking multiple boxes, ticking the method used that week as opposed to the usual method and confusion over the meaning of the category 'car or van – pool, sharing driving'. The 1991 form dispensed with this category and this may explain why the gross error rate was reduced from 9% in 1981 to 8.1% in 1991.

Ethnic group

The ethnic group question was newly introduced in the 1991 Census. After extensive

testing in the field, it was decided to use a question which gave form fillers nine possible categories from which to choose, two of which asked for more detailed information to be supplied. The number of ethnic minority households in the CVS sample was not sufficient to justify individual analysis of all nine categories and so four aggregate codes were created: white, black (combining black Caribbean, black African and black 'other'), Indian sub-continent (Indian, Pakistani and Bangladeshi) and other (Chinese and 'any other ethnic group'). The gross error rate was only 0.8 per cent. However, this figure should be treated with caution given that the vast majority of answers were in just one category (white). If those who answered 'white' in both the Census and CVS are excluded, the gross error rate was 13.2 per cent. It was found that 21 per cent of those coded as 'other' in the Census either described themselves as white or in one of the black categories in the CVS. Conversely, 9 per cent of those coded 'black' in the Census described themselves as 'other' in the CVS. Overall, 6.1 per cent of people in households who replied in both the Census and CVS were in the non-white ethnic group in the CVS, compared with 5.8 per cent of the same people according to their replies in the Census (Heady et al, 1996).

3.4. Incomplete Coverage

Two types of resident are missing from the SARs. This section is not concerned with the effects of sampling, which have been described earlier, but with the completeness of the census itself. The SARs were drawn from the fully coded set of census records returned by households and institutions. There are two main categories of residents missing from this set of records:

Imputed absent households

Census records representing 869,000 residents in Great Britain (1.6 per cent) have been imputed by the census offices by using the enumerators' estimate of the number of residents in absent households and then copying characteristics from geographically adjacent households who returned late census forms. These records were used in compiling the 100 per cent tabular census output, but are excluded from the 1991 SARs, which are drawn from the 10 per cent sample. This imputation procedure was followed wherever an enumerator felt that a housing space contained residents but a) a household absent on census night did not return a form under the voluntary arrangements; or b) the enumerator could make no contact at all; or c) residents refused to complete and return a form.

Other residents missed by the census

Imputation of absent residents 'captured' less than half of the number estimated not to have been enumerated in the census. In some cases residents were not included on the census forms that were returned. In other cases whole households were missed by enumerators who had difficulty enumerating, for example, those living in converted and multi-occupied properties. The number of residents not included in the 100 per cent output, that is neither enumerated nor imputed, was estimated to have been 1.2 million in Great Britain (2.1 per cent).

[Editor: Since the 2001 census the assumptions above have been challenged. The following text is taken a letter from Len Cook, the then National Statistician, quoted in Hansard at
https://publications.parliament.uk/pa/ld200203/ldhansrd/vo021205/text/21205w03.htm
<last accessed 1/5/17>:
'The UK has 800,000 fewer young men than previously thought. This pattern was originally identified in the 1991 census, but given a lack of confidence in the follow up

survey for that census, the numbers were revised to restore the predicted pattern. In 2001, the pattern has been confirmed and validated by the one number census. We will revise population estimates back to 1982. The critical factor appears to be emigration. The International Passenger Survey works well, but it captures travellers' intentions at the time of departure. These may be prone to change once people are abroad, particularly among young men with few ties at home.']

A small number of individuals are excluded from the household SAR for confidentiality reasons: those where the number of persons in the household is twelve or more. These comprise 28 (0.013 per cent) households, approximately 0.06 per cent of residents in households.

3.5. Allowing for incomplete census coverage in analysis using the SARs

Often, the user will wish to make an inference from the SARs about conditions of the full population on Census night in 1991. Those missing from the Census data from which the SARs were drawn may have distinctive characteristics. To enable users to compensate for those missing in the SARs (wholly absent imputed households) plus those missed from the Census, weights are being added to the SARs which allow adjustment to the mid 1991 population estimates. These are specific to age, sex and SAR area are available as a derived variable (POPWGHT). It is important to note that population estimates are based on residents and therefore visitors should be excluded from analysis when applying population weights.

For example, to get an accurate age/sex profile of every SAR area, variables should first be weighted by POPWGHT. The population weights could be used in a similar way to obtain estimates of individuals in other sub- populations such as ethnic groups. However, when using the population weight only age, sex and SAR area are taken into consideration. Weighting assumes that the characteristics of the imputed and missing population are the same as those of the sampled population. In general weights should be used with caution when looking at variables which have small cell sizes because very small groups may be disproportionately boosted by a large weighting factor. The method of making such adjustments, and the magnitude of the impact of census undercount on some simple census indicators is described more fully in Simpson (1993) 'Measuring and coping with local under-enumeration' Paper presented to the 'Research on the 1991 Census' Conference, Newcastle upon Tyne September 1993.

4. Selecting a population base

Unlike in the aggregate data (called Small Area Statistics (SAS) and Local Base Statistics (LBS)), population bases and units of analysis within the SARs are not predetermined. This means that users must be careful to select the appropriate population base and unit of analysis before any analysis is carried out. The options are different for the individual and household SARs.

4.1. Individual Files

Users may choose between the following population bases and should select on the variables RESIDSTA/RESIDNI and CESTSTAT/CESTATNI as appropriate.

In households:

- Present residents
- Absent residents visitors
- Visitors

In communal establishments:

- Residents (non staff)
- Visitors
- Residents (staff)

Because visitors, by definition, may be absent residents at another address, inclusion of both groups may cause double counting of some people. Normally the user should select *either* residents *or* visitors. Students, for example, are instructed on the Census form to record themselves as visitors if enumerated at their term time address. However, they should also be recorded as absent residents at their parental address. If by chance, both the student record at term time and vacation address were selected for the sample, the student would be double counted.

Users should also be careful to take note of the population bases for particular variables. For example, if looking at occupation, industry, hours of work, socio-economic group or social class, each category includes people who have been in paid work in the previous ten years, not just those who were in paid work at the time of the Census. To achieve comparability with other sources (e.g. the LBS) it may be necessary to select only residents in employment.

4.2. Household files

The household file does not contain communal establishments, but otherwise the same guidelines apply. In addition, however, users should be careful to select the correct unit of analysis from:

- Individuals
- Families (not available in the Northern Ireland Household SAR)
- Households

Using SPSS or similar package, a table extracted from the household file will give counts of persons in households. To obtain data at the level of household, it is necessary to select one person per household. For example, if we were to produce a table of the households' car access by tenure we could select one person per household by selecting those cases where PNUM=1 (which would return the first person number in each household).

5. Geography of the SARs

5.1. The 2% Individual SAR (Great Britain)

As noted, empirical work and comparison with public use datasets from other countries showed that a sensible geographical level for release of the individual SAR was areas equivalent to large local authority districts. Each SAR area must have a population size of at least 120,000 based on the mid-1989 estimates. Either as aggregations or in their own right, these areas allow all non-metropolitan counties in England and Wales, most Scottish regions, all metropolitan districts and all London boroughs (with the exception of the City of London) to be identified. Local authority districts with populations of less than 120,000 were grouped to form larger areas on the basis of four rules. First, the integrity of county/Scottish region geography was maintained where possible. Secondly, districts which achieved the minimum 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. Fourthly, if a choice remained once the above criteria had been met, areas were grouped on the basis of their apparent social and historical similarity. Figure 1 shows the resulting 278 districts.

The SAR areas (with two exceptions in Scotland) can be aggregated to correspond with counties in England and Wales and regions in Scotland, and this aggregation is included as a standard derived variable within the individual file. County areas can in turn be aggregated up to SAR region level and standard region level.

5.2. The 1% Household SAR (Great Britain)

As a protective measure, given the greater risks of disclosure from the household SAR, the lowest geographical detail used for the 1% household SAR is the Registrar General's Standard Regions, plus Wales and Scotland. The only exception is the South East which has been split into Inner London, Outer London, and the rest. The full list is:

01 North
02 Yorkshire and Humberside
03 East Midlands
04 East Anglia
05 Inner London
06 Outer London
07 Rest of South East
08 South West
09 West Midlands
10 North West
11 Wales
12 Scotland

5.3. The 2% Individual SAR (Northern Ireland)

The population thresholds applied to the SAR areas for Great Britain (areas with populations in excess of 120,000 people) apply to the Northern Ireland SARs. Consequently, the 2% Individual SAR allows for ten SAR areas, based on amalgamations of co-terminus District Councils with similar characteristics:

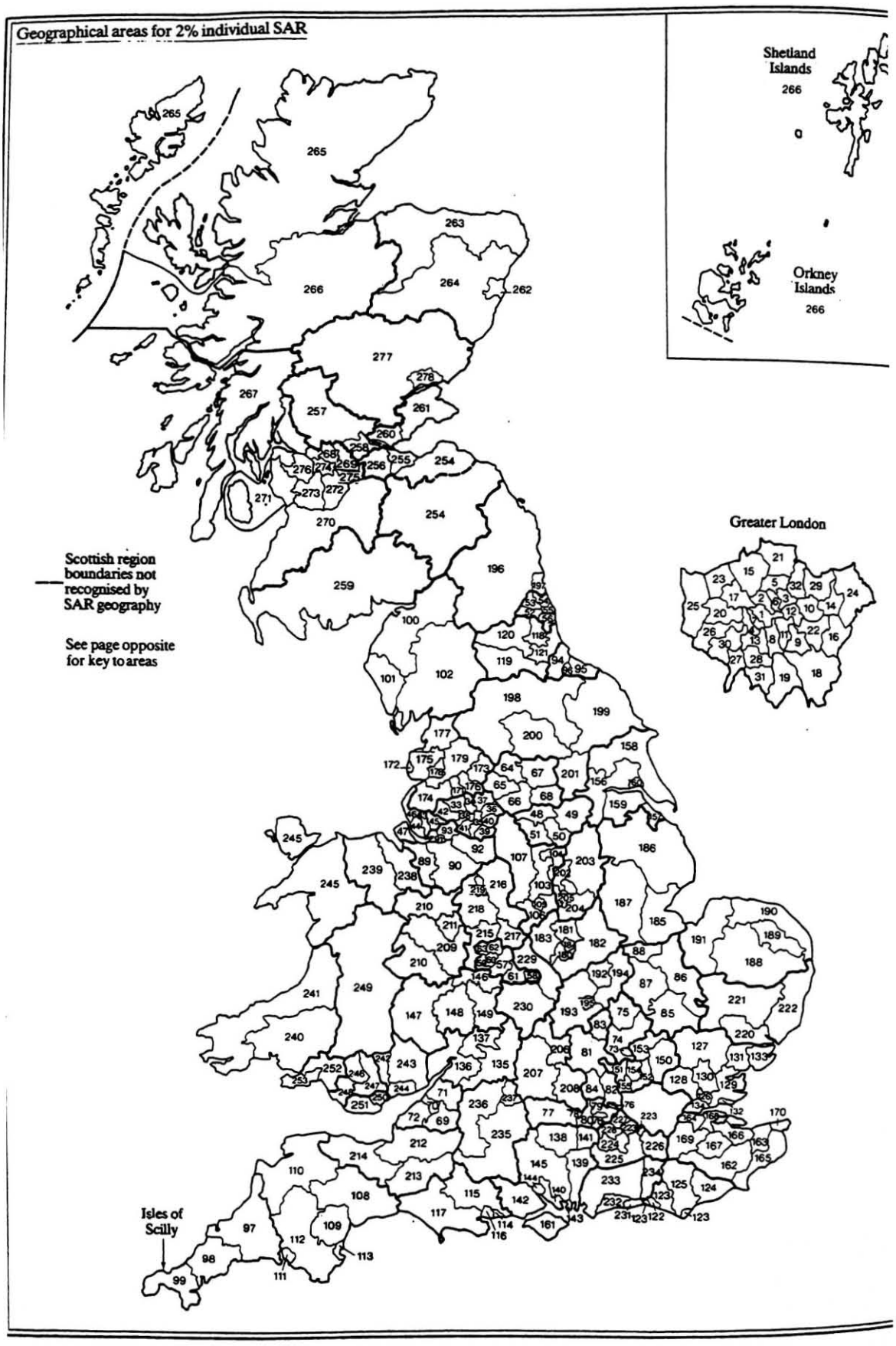
01 Belfast
02 Ards, Castlereagh, North Down
03 Down, Lisburn
04 Carrickfergus, Larne, Newtownabbey
05 Antrim, Ballymena, Balleymoney
06 Coleraine, Cookstown, Margherafelt, Moyle
07 Armagh, Newry & Mourne
08 Banbridge, Craigavon, Dungavvon
09 Derry, Limavady
10 Fermanagh, Omagh, Strabane

5.4. The 1% Household SAR (Northern Ireland)

For the Northern Ireland 1% Household SAR, the geographical level at which the data are released is that of Northern Ireland itself.

5.5. SAR geography and other levels of Census geography

SAR areas in the Individual SAR are made up of aggregations of Local Government Districts (LGD) in Great Britain and of District Councils in Northern Ireland. These districts correspond to the geography of the SAS/LBS.


Geographical areas for 2% individual SAR
Shetland Islands
266
Orkney Islands
266
Greater London
Scottish region boundaries not recognised by SAR geography
See page opposite for key to areas
Isles of Scilly

Key to Map – Geographical areas for two per cent individual SAR

Local authorities/London boroughs

Inner London

1 City of London; City of Westminister
2 Camden
3 Hackney
4 Hammersmith and Fulham
5 Haringey
6 Islington
7 Kensington and Chelsea
8 Lambeth
9 Lewisham
10 Newham
11 Southwark
12 Tower Hamlets
13 Wandsworth

Outer London

14 Barking and Dagenham
15 Barnet
16 Bexley
17 Brent
18 Bromley
19 Croydon
20 Ealing
21 Enfield
22 Greenwich
23 Harrow
24 Havering
25 Hillingdon
26 Hounslow
27 Kingston upon Thames
28 Merton
29 Redbridge
30 Richmond upon Thames
31 Sutton
32 Waltham Forest

Greater Manchester

33 Bolton
34 Bury
35 Manchester
36 Oldham
37 Rochdale
38 Salford
39 Stockport
40 Tameside
41 Trafford
42 Wigan

Merseyside

43 Knowsley
44 Liverpool
45 St Helens
46 Sefton
47 Wirral

South Yorkshire

48 Barnsley
49 Doncaster
50 Rotherham
51 Sheffield

Tyne and Wear

52 Gateshead
53 Newcastle upon Tyne
54 North Tyneside
55 South Tyneside
56 Sunderland

West Midlands

57 Birmingham
58 Coventry
59 Dudley
60 Sandwell
61 Solihull
62 Walsall
63 Wolverhampton

West Yorkshire

64 Bradford
65 Calderdale
66 Kirklees
67 Leeds
68 Wakefield

Avon

69 Bath; Kingswood; Wansdyke
70 Bristol
71 Northavon
72 Woodspring

Bedfordshire

73 Luton
74 Mid Bedfordshire; South Bedfordshire
75 North Bedfordshire

Berkshire

76 Bracknell Forest; Slough
77 Newbury
78 Reading
79 Windsor & Maidenhead
80 Wokingham

Buckinghamshire

81 Aylesbury Vale
82 Chiltern
83 Milton Keynes
84 Wycombe

Cambridgeshire

85 Cambridge; South Cambridgeshire
86 East Cambridgeshire; Fenland
87 Huntingdonshire
88 Peterborough

Cheshire

89 Chester; Ellesmere Port and Neston
90 Congleton; Crewe and Nantwich; Vale Royal
91 Halton
92 Macclesfield
93 Warrington

Cleveland

94 Hartlepool; Stockton-on-Tees
95 Langbaurgh-on-Tees
96 Middlesbrough

Cornwall & Isles of Scilly

97 Caradon; North Cornwall
98 Carrick; Restormel
99 Kerrier; Penwith; Isles of Scilly

Cumbria

100 Allerdale; Carlisle
101 Barrow-in-Furness Copeland
102 Eden; South Lakeland

Derbyshire

103 Amber Valley; North East Derbyshire
104 Bolsover; Chesterfield
105 Derby
106 Erewash; South Derbyshire
107 High Peak; The Derbyshire Dales

Devon

108 East Devon; Mid Devon
109 Exeter; Teignbridge
110 North Devon; Torridge
111 Plymouth
112 South Ham; West Devon
113 Torbay

Dorset

114 Bournemouth
115 Christchurch; East Dorset North Dorset
116 Poole
117 Purbeck; West Dorset; Weymouth and Portland

Durham

118 Chester-le-Street; Durham
119 Darlington; Teesdale
120 Derwentside; Wear Valley
121 Easington; Sedgefield

East Sussex

122 Brighton
123 Eastbourne; Hove; Lewes
124 Hastings; Rother
125 Wealdon

Essex

126 Basildon
127 Braintree; Uttlesford
128 Brentwood; Epping Forest; Harlow
129 Castle Point; Maldon; Rochford
130 Chelmsford
131 Colchester
132 Southend-on-Sea
133 Tendring
134 Thurrock

Gloucestershire

135 Cheltenham; Cotswold
136 Forest of Dean; Stroud
137 Gloucester; Tewkes'bury

Hampshire

138 Basingstoke & Deane
139 East Hampshire; Havant
140 Eastleigh; Fareham; Gosport
141 Hart; Rushmoor
142 New Forest
143 Portsmouth
144 Southampton
145 Test Valley; Winchester

Hereford and Worcester

146 Bromsgrove; Wyre Forest
147 Hereford; Leominster; South Herefordshire
148 Malvern Hills; Worcester
149 Redditch; Wychavon

Hertfordshire

150 Broxbourne; East Hertfordshire
151 Dacorum
152 Hertsmere; Welwyn Hatfield
153 North Hertfordshire; Stevenage
154 St Albans
155 Three Rivers; Watford

Humberside

156 Beverley; Boothferry
157 Cleethorpes; Great Grimsby
158 East Yorkshire; Holderness
159 Glanford; Scunthorpe
160 Kingston-upon-Hull

Isle of Wight

161 Medina; South Wight

Kent
162 Ashford; Tunbridge Wells
163 Canterbury
164 Dartford; Gravesham
165 Dover; Shepway
166 Gillingham; Swale
167 Maidstone
168 Rochester upon Medway
169 Sevenoaks; Tonbridge & Malling
170 Thanet

Lancashire
171 Blackburn
172 Blackpool
173 Burnley; Pendle
174 Chorley; West Lancashire
175 Fylde; Wyre
176 Hyndburn; Rossendale
177 Lancaster
178 Preston
179 Ribble Valley; South Ribble

Leicestershire
180 Blaby; Oadby & Wigston
181 Charnwood
182 Harborough; Melton; Rutland
183 Hinckley & Bosworth; North West Leicestershire
184 Leicester

Lincolnshire
185 Boston; South Holland
186 East Lindsey; Lincoln; West Lindsey
187 North Kesteven; South Kesteven

Norfolk
188 Breckland; South Norfolk
189 Broadland; Norwich
190 Great Yarmouth
191 Kings Lynn & West Norfolk

Northamptonshire
192 Corby; Kettering
193 Daventry; South Northamptonshire
194 East Northamptonshire; Wellingborough
195 Northampton

Northumberland
196 Alnwick; Berwick-upon-Tweed; Castle Morpeth; Tynedale
197 Blyth Valley, Wansbeck

North Yorkshire
198 Craven; Hambleton; Richmondshire
199 Ryedale; Scarborough
200 Harrogate
201 Selby; York

Nottinghamshire
202 Ashfield; Mansfield
203 Bassetlaw; Newark & Sherwood
204 Broxtowe; Gedling; Rushcliffe
205 Nottingham

Oxfordshire
206 Cherwell
207 Oxford; Vale of White Horse; West Oxfordshire
208 South Oxfordshire

Shropshire
209 Bridgnorth; Shrewsbury & Atcham
210 North Shropshire; Oswestry; South Shropshire
211 The Wrekin

Somerset
212 Mendip; Sedgemoor
213 South Somerset
214 Taunton Deane; West Somerset

Staffordshire
215 Cannock Chase; South Staffordshire
216 East Staffordshire; Staffordshire Moorlands
217 Lichfield; Tamworth
218 Newcastle-under-Lyme; Stafford
219 Stoke-on-Trent

Suffolk
220 Babergh; Ipswich
221 Forest Heath; Mid Suffolk; St Edmundsbury
222 Suffolk Coastal; Waveney

Surrey
223 Elmbridge; Epsom & Ewell
224 Guildford
225 Mole Valley; Waverley;
226 Reigate & Banstead; Tandridge
227 Runnymede; Spelthorne
228 Surrey Heath; Woking

Warwickshire
229 North Warwickshire; Nuneaton & Bedworth; Rugby
230 Stratford-on-Avon; Warwick

West Sussex
231 Adur; Worthing
232 Arun
233 Chichester; Horsham
234 Crawley; Mid Sussex

Wiltshire
235 Kennet; Salisbury
236 North Wiltshire; West Wiltshire
237 Thamesdown

Clwyd
238 Alyn & Deeside; Delyn; Wrexham Maelor
239 Colwyn; Glyndwr; Rhuddlan

Dyfed
240 Carmarthen; Dinefwr; Llanelli
241 Ceredigion; Preseli Pembrokeshire; South Pembrokeshire

Gwent
242 Blaenau Gwent; Islywn
243 Monmouth; Torfean
244 Newport

Gwynedd
245 Aberconwy; Arfon; Dwyfor; Meirionnydd; Ynys Mon - Isle of Anglesey

Mid Glamorgan
246 Cynon Valley; Rhondda
247 Merthyr Tydfil; Rhymney Valley; Taff-Ely
248 Ogwr

Powys
249 Brecknock; Montgomeryshire; Radnorshire

South Glamorgan
250 Cardiff
251 Vale of Glamorgan

West Glamorgan
252 Lliw Valley; Neath; Port Talbot
253 Swansea

Borders and Lothian
254 Berwickshire; East Lothian; Ettrick & Lauderdale; Mid Lothian; Roxburgh; Tweeddale
255 Edinburgh City
256 West Lothian

Central
257 Clackmannan; Stirling
258 Falkirk

Dumfries & Galloway
259 Annandale & Eskdale; Nithsdale; Stewarty; Wigtown

Fife
260 Dunfermline
261 Kirkcaldy; North East Fife

Grampian
262 Aberdeen City
263 Banff & Buchan; Moray
264 Gordon; Kincardine & Deeside

Highland and Islands Areas
265 Caithness; Sutherland; Ross & Cromarty; Skye & Lochalsh; Western Isles
266 Badenoch & Strathspey; Inverness; Lochaber; Nairn; Orkney Islands; Shetland Islands

Strathclyde
267 Argyll & Bute; Dumbarton; Inverclyde
268 Bearsden & Milngavie; Clydebank; Strathkelvin
269 Cumbernauld & Kilsyth; Monklands
270 Clydesdale; Cumnock & Doon Valley; Kyle & Carrick
271 Cunninghame
272 East Kilbride; Hamilton
273 Eastwood; Kilmarnock & Loudon
274 Glasgow City
275 Motherwell
276 Renfrew

Tayside
277 Angus; Perth & Kinross
278 Dundee City

6. References

Clark A.M. 1992. 1991 Census: data collection, *Population Trends*, 70, 22–27.

Cole, K. (1993) The 1991 Local Base and Small Area Statistics, in Dale A and Marsh, C, *The Census Users Guide* London, HMSO

Dale, A and Marsh, C (1993) The Census Users Guide London, HMSO

Heady, Patrick. (1996) "Census validation methods, with special reference to identifying which dwellings are occupied and to capture/recapture estimation." *Looking towards the 2001 census: papers given at a joint BSPS and RSS conference held on April 6th 1995 at City University*. Vol. 46. OPCS

Marsh, C., Arber, S., Wrigley, N., Rhind, D. and Bulmer, M. (1988) Research policy and review 23: The view of academic social scientists on the 1991 UK Census of Population: a report of the Economic and Social Research Council Working Group, *Environment and Planning A, 20*, 851-889

Middleton, E (1993) Problems of harmonising UK-wi de Samples of Anonymised Records from the 1991 Census, *Paper presented to the Annual Conference of IASSIST/IFDO* Edinburgh, May 1993

Mills, I. and Teague, A. (1991) Editing and imputing data for the 1991 Census, *Population Trends 64,* pp30-37

OPCS and GRO(S) (1992) *1991 Census Definitions (Great Britain)* London: HMSO

Simpson, S. (1993) "Measuring and coping with local under-enumeration" *Paper presented to the "Research on the 1991 Census" Conference*, Newcastle upon Tyne, September 1993

Wiggins, R (1993) "The validation of Census data: (1) Post-enumeration survey approaches" in Dale A and Marsh C (eds) *The Census Users Guide* London, HMSO