

## 4. 2001 SARs

### 4.1 Introduction

The 1991 SARs have been widely used for a range of high quality research. The key findings and publications are available on the web site.

Following extensive consultation with users a request was submitted to ONS in September 2001 for three datasets: a 3 per cent Individual SAR, a 1 per cent Household SAR and a 5 per cent Small Area Microdata file (SAM). The latter was particularly requested by geographers who were concerned to obtain more geographical detail at the expense of individual information. The full details of the justification for the request and the specifications can be downloaded from [www.ccsr.ac.uk/sars/2001/request/sarequest.pdf](http://www.ccsr.ac.uk/sars/2001/request/sarequest.pdf).

In brief the request sought to:

- reduce the population threshold for SAR areas in the Individual SAR from 120k to 90k
- increase the sample size for the Individual SAR from 2 per cent to 3 per cent
- include additional detail in some of the variables, for example ethnic group, family type and professional qualifications, to reflect changes in the information collected in 2001
- add extra variables (to reflect the new questions asked in the 2001 Census).

However, increased concerns about confidentiality of microdata has resulted in the Individual and Household files being significantly less detailed than the original request and also less detailed than the 1991 SARs. In response to this ONS have established Controlled Access Microdata Samples (CAMS), which are only accessible within a safe setting in the statistical offices of the UK.

Both concern over under-enumeration and the importance of obtaining accurate population estimates and also concern over confidentiality have had marked influences on the 2001 census. The former has resulted in the 'One Number Census' and the latter has influenced both the outputs available and also their timing. Both are discussed in the sections below.

#### 4.1.1 The One Number Census

The 2001 Census aimed to maximise coverage and to make an accurate estimate of the people missed. The 1991 Census was thought to have had a substantially larger under-count than in previous censuses with about 2 per cent of the population of GB missed entirely and a further 1.6 per cent for whom records were imputed.

The One Number Census was designed to produce figures from the 2001 Census that are adjusted for under-enumeration and which are consistent across all forms of output and at the smallest geographical area. The term 'One Number Census' indicates a departure from the 1991 Census where preliminary figures from the census count were published and then later figures, adjusted for under-enumeration, were published. The One Number Census approach makes all adjustments as part of the census processing. Thus the One Number Census results in a database of the complete population for the UK from which all census outputs – including the SARs - are drawn.

The key stages of the ONC can be summarised as follows:

- a) A Census Coverage Survey (CCS), undertaken independently of the Census, was designed to establish the coverage of the 2001 Census. For the CCS, the UK was divided into one hundred and twelve areas, each with a population of about 500,000. These areas are known as design groups and are made up of whole LADs or groups of smaller LADs. The CCS took place in all of these design groups.
- b) The CCS records are matched with those from the Census using a combination of automated and clerical matching.
- c) Populations for each design group, by age and sex, are estimated using a combination of standard estimation techniques.
- d) Small area estimation techniques are used to estimate Local Authority District populations by age and sex.
- e) Households and individuals estimated to be missed by the Census are imputed to produce a fully adjusted Census database.
- f) All ONC population estimates are quality assured using demographic analysis and aggregate level administrative data.

More detail on the One Number Census is available at [www.statistics.gov.uk/nsbase/census2001/pdfs/oncguide.pdf](http://www.statistics.gov.uk/nsbase/census2001/pdfs/oncguide.pdf).

#### **4.1.2 Licensed Individual SAR**

The Licensed Individual SAR is safe data that is available to registered users for analysis outside ONS. It is a 3 per cent sample and contains 1,843,530 individuals and includes information on age, gender, ethnicity, health, employment status, housing, amenities, family type, geography, social class, education, distance to work, workplace, hours worked and migration. The 3 per cent sample is an increase by comparison with 2 per cent in 1991.

In addition, the ONS have added occupational coding, not available in the census tables, for individuals aged 16-65 who last worked more than 5 years ago but less than ten years ago and for those aged 65-74 who were not currently working at the census but who had worked in the previous ten years. A full list of variables is available at [www.ccsr.ac.uk/sars/2001/individ/variables/index.html](http://www.ccsr.ac.uk/sars/2001/individ/variables/index.html).

The lowest level of geography is the Government Office Region, although Inner and Outer London are separately identified. This represents a significant reduction by comparison with the 1991 where large Local Authorities (population 120K and over) were separately identified. A quick comparison between 1991 and 2001 SARs can be found in section 5.5 of this document.

The data are available online to registered users. There is no charge for academic use. Public sector bodies can obtain the data free of charge and the business sector are charged £1000 per file. The licence will entitle the organisation to receive the data which may then be accessed by ten people, all of whom will have to sign a user undertaking. Details of registration and access are available at [www.ccsr.ac.uk/sars/access/](http://www.ccsr.ac.uk/sars/access/).

#### 4.1.3 Special Licence Household Sample of Anonymised Records (SL-HSAR)

The Special License Household SAR (SL-HSAR) is a 1 per cent sample of households and all those individuals in those households from the 2001 Census. It is a hierarchical file allowing linkages to be made between individuals within families and households. The Special License Household SAR contains information on age, gender, ethnicity, marital status, social class, education and employment status. It also includes household level variables, e.g. housing tenure and number of cars. A number of derived variables have been added, for example, the number of full time earners in a household or the age of the youngest dependent child in a household.

The 2001 Special License Household SAR is available for England and Wales only. The Special License Household SAR represents a one percent sample of all households drawn from the 2001 Census. It comprises 225436 household records and 525715 individual records. Individual records are available only for households with 11 or fewer residents. Household records include a small number of empty households. For households of 12 or over, only household level variables are available – there are no individual records. To protect confidentiality age has been grouped into two-year bands and there is no geographical breakdown available. There has been a small amount of perturbation to protect confidentiality.

Whilst the actual numbers of individuals and households are relatively small, large households are not randomly distributed in the population. For example, the loss of this information would disproportionately affect

Pakistani and Bangladeshi ethnic groups and would bias estimates of overcrowding and various forms of deprivation.

This file is only available under an Office for National Statistics (ONS) special licence, via the UK Data Archive. Users have to agree to keep the data under secure conditions and institutions are responsible for ensuring these conditions are met. For more information see the following link

<http://www.data-archive.ac.uk/orderingdata/specialLicence.asp>

The specification and codebook of the Special License Household SAR can be found in the following link.

<http://www.ccsr.ac.uk/sars/2001/hhold/>

The sample excludes those in communal establishments. It includes households with dummy forms and also 'students living-away' who provide very limited individual information because they should be fully enumerated at their usual term-time residence.

The variable 'popbase' allows users to identify these categories and select their required population base. The file was sampled from the 'one-number census' database. This includes data which was imputed for non-respondent individuals or households. Imputed cases can be identified using the "oncperim" variable.

#### 4.1.4 Licensed Small Area Microdata

The Small Area Microdata (SAM) is a 5 per cent sample of individuals from the 2001 Census with local authority as the lowest level of geography. Because of confidentiality concerns there will be less individual detail than on the Individual SAR. The strength of the data lies in the more detailed geography. The SAM is for all countries of the UK, with 2.96 million cases. Local Authority is the lowest level of geography for England and Wales, Council Areas for Scotland and Parliamentary Constituencies for Northern Ireland. The Scilly Isles have been merged with Penwith and the City of London with Westminster. For Scotland, Orkney and Shetland are merged into one area. All other areas are identified. The file contains less individual detail than the Individual SAR:

- Age in 13 categories
- Economic activity in 4 categories
- Ethnic group in 13 categories in England and Wales, 8 in Scotland and 2 in Northern Ireland
- NS-SEC in 8 categories

Further details are at <http://www.ccsr.ac.uk/sars/2001/sam/>

#### 4.1.5 Controlled Access Microdata Sample (Individual and Household)

In recognition of the reduction in detail in the Individual SAR, a more detailed dataset is available in a safe setting. Initially the data is available (version 1) with full ethnic group information (16 categories), age in single years to 95, SOC minor and 60 categories of industry. This file will also contain details of local authority and the Index of Multiple deprivation. A codebook and specifications for the individual data are currently available at [www.ccsr.ac.uk/sars/2001/indiv-cams/codebook/](http://www.ccsr.ac.uk/sars/2001/indiv-cams/codebook/).

A Household Controlled Access Microdata Sample (Household CAMS) is also available and includes data for Scotland and Northern Ireland. A codebook is available at [www.ccsr.ac.uk/sars/2001/hhold-cams/codebook/index.html](http://www.ccsr.ac.uk/sars/2001/hhold-cams/codebook/index.html).

The files are available for research use only and applications must be made to the Census Research Access Board at the Office for National Statistics. Details of access are available for the Individual CAMS at [www.ccsr.ac.uk/sars/2001/indiv-cams/access/](http://www.ccsr.ac.uk/sars/2001/indiv-cams/access/) and for the household CAMS at [www.ccsr.ac.uk/sars/2001/hhold-cams/](http://www.ccsr.ac.uk/sars/2001/hhold-cams/).

A quick comparison of the different specifications for all variants of the 2001 SAR files can be found **at** [www.ccsr.ac.uk/sars/guide/2001comparison.pdf](http://www.ccsr.ac.uk/sars/guide/2001comparison.pdf)

#### 4.1.6 CAMS Test File

The CAMS test file is a sub-sample of 298,912 cases from the Individual Controlled Access Microdata Sample from the 2001 Census. Variables have been perturbed to ensure that no sample members can be identified. Perturbation has retained the correct distribution of each variable but the relationships between variables will not give expected results. The test file can therefore be used to develop and test syntax for analyses before going to use the CAMS at ONS. It cannot be used to test exploratory analyses with any reliability, nor will you be able to test statistical procedures which are dependent on distributions. It is *not* suitable for research purposes and has been provided *only* as a dummy dataset for preparing syntax in advance of using the CAMS file.

The protection comes from perturbing multiple variables per person with a high probability of change and by providing no indication of whether a variable value has been perturbed.

#### **Availability of the synthetic test dataset**

The CAMS test file is available for download under the standard End User Licence as SPSS and STATA files. For further information see <http://www.ccsr.ac.uk/sars/2001/test-cams/>

## Coverage

The CAMS test file only covers:

- England and Wales
- Individuals in private households.

and excludes:

- Students living away from their parental home
- households of size 6 or more

These limitations do not apply to the full CAMS files. Other differences apply in terms of variable availability. Users should consult the codebook for the full CAMS file prior to applying to use the data or preparing syntax.

### **The following variables are in the full CAMS file but not in the test file**

cestatux	Status in communal establishment (extended)
cetypews	Type of Communal Establishment, England Wales and Scotland
cetypn	Type of Communal Establishment, Northern Ireland
cobpuk	country of birth
combgn	community background, religion or religion brought up in
distmov0	distance of move for migrants - distance in bands
dstwrk0	Distance to Work (Including Study in Scotland)
ethnx	Ethnic Group for Northern Ireland
ethsx	Ethnic Group for Scotland
furn	Accommodation Furnished- (Scotland Only)
gaelread	Whether reads Gaelic (Scotland only)
gaelspk	Whether speaks Gaelic (Scotland only)
gaelstnd	Whether understand Gaelic (Scotland only)
gaelwrit	Whether Writes Gaelic (Scotland only)
irisread	Whether reads Irish (NI only)
irisstk	Whether Speaks Irish (NI only)
irisstnd	Whether Understands Irish (NI only)
iriswrit	Whether Writes Irish (NI only)
isco	International standard classification of occupations
qualvs	(Scotland only) - Level of Highest Qualifications (Aged 16 to 74)
relgn	Religion (Northern Ireland) belongs to/brought up in
relgs1	Religion belongs to (SCOTLAND)
relgs2	Religion brought up in (SCOTLAND)
roomsflr	Rooms used by Household on More than 1 Floor (NI only)
socunit	Occupations (SOC 2000 unit)
tenurnsn	Tenure of Accommodation (Scotland and Northern Ireland only)
urbrurs	Urban/Rural - Scotland
<b>The following variables are in the test file but not in the full</b>	

<b>CAMS file</b>	
dcobuk	Derived country of birth
ddistmov	Distance moved for migrants-derived banded variable

#### *4.2 Sampling the 2001 SARs*

The 2001 SARs were sampled from the One Number Census database for the entire UK. Thus individuals and households imputed as part of the ONC are included in the SARs.

Sampling was done within each of the 112 ONC design groups for the UK. Design groups represent groups of LADs, each with a population of approximately half a million people, which form an essential part of the ONC process. The CCS sampling was applied within each of the design groups to enable accurate direct estimates of under enumeration for 37 age-sex groups at the design group level.

There are 112 design groups in the UK (101 in England and Wales, eight in Scotland and three in Northern Ireland).

The sampling scheme for the household SAR is a stratified simple random sampling, where the strata are EDs (there is a very minor departure from simple random sampling in the sense that the sample size within EDs may vary randomly between two adjacent integers because of the random start from 1 to 10). Unlike 1991, there is no stratification within EDs. Random sampling is applied within each ED.

The sampling scheme for the individual SAR follows the 1991 approach of drawing from the population excluding the household sample. Stratification is again by ED. The Individual SAR sampled both private and communal persons, unlike the household SAR which only sampled only households.

As for the 1991 sample, 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.

There is no overlap between any of the SAR files. All records have been scrambled before release to ensure they cannot be identified geographically.

### *4.3 Differences between countries of the UK*

In the UK there are separate statistical offices for England and Wales, Scotland and Northern Ireland and the Registrar General in each office has responsibility for the content and conduct of the census of population for that country. Therefore each office makes its own decisions over the questions to be asked and the methods to be used to produce outputs. Despite this there is a good deal of co-ordination of topics and questions between the Office for National Statistics (England and Wales), the General Register Office for Scotland and the Northern Ireland Statistics and Research Agency. Whilst the objective has been to ask the same questions across the United Kingdom, there are a number of differences resulting from specific requirements in each country. These differences have an effect on the classifications and outputs available for each country of the UK.

The UK 2001 Census Definitions volume provides the main source of information on the methods used to conduct the 2001 census and the differences between countries. A few differences of significance to SAR users are highlighted below.

#### **4.3.1 Question differences**

##### *Household Questions*

**Number of floors.** Asked in Northern Ireland only.

**Landlord.** This question is the same for all parts of the UK, except that the response categories use local terminology.

**Furnished/Unfurnished.** Asked in Scotland only.

##### *Person Questions*

**Ethnic group.** There are differences in this question between England and Wales, Scotland and Northern Ireland, resulting in separate output classifications.

**Religion.** There are differences in this question between England and Wales, Scotland and Northern Ireland, resulting in separate output classifications. There are two questions in Scotland and Northern Ireland (current religion and religion of upbringing), only one in England and Wales.

**Language.** There are specific questions on Celtic languages asked in Wales (Welsh), Scotland (Gaelic) and Northern Ireland (Irish). England has no question on Celtic language.

**Qualifications.** There are differences in this question between England and Wales, Scotland and Northern Ireland. In England and Wales there is an additional question asking for information on teaching, medical, nursing and dental professional qualifications. In Scotland there is one tick box response category asking if people have 'professional qualifications (for example, teaching, accountancy)'. In Northern Ireland, there is no question asking for information on professional qualifications.

**Travel destination.** In England, Wales and Northern Ireland the travel destination question refers only to place of work and is asked of people aged 16 to 74 in a job the week before Census. In Scotland, the travel destination question is asked of all people and relates to the address travelled to for the main job or course of study (including school).

**Method of travel.** In England, Wales and Northern Ireland the method of travel relates to travel to the main place of work. In Scotland, method of travel relates to travel to the main place of work or study. There are minor differences in the response categories across countries. For more information about question differences between 1991 and 2001 across the UK countries see the following document.

[http://www.statistics.gov.uk/census2001/pdfs/class\\_sections1\\_3.pdf](http://www.statistics.gov.uk/census2001/pdfs/class_sections1_3.pdf)

#### **4.3.2 Variables in the Licensed Samples of Anonymised Records**

For confidentiality reasons ethnic group is only available in Northern Ireland as a dichotomy: White; other ethnic group. For a complete list of variables in the Licensed Individual SARs see

<http://www.ccsr.ac.uk/sars/2001/indiv/variables/>

#### **4.3.3 Classifications used in the Samples of Anonymised Records**

ONS has produced a classifications document that provides comprehensive information on 2001 Census definitions, concepts and classifications. The document has been split into four files, accessible from [www.statistics.gov.uk/census2001/outputclassifications.asp](http://www.statistics.gov.uk/census2001/outputclassifications.asp).

There are four household indicator variables, associated with different dimensions of deprivation: housing, health, employment and education. These classifications are only available for England and Wales and Northern Ireland.

The Census results include a table which measures deprivation according to four 'dimensions' of deprivation:

Education: No member of the household aged 16 to pensionable age has at least 5 GCSEs (grade A-C) or equivalent AND no member of the household aged 16-18 is in full-time education.

Health and disability: Any member of the household has general health 'not good' in the year before Census or has a limiting long term illness.

Housing: The household's accommodation is either overcrowded (occupancy indicator is -1 or less), OR is in a shared dwelling OR does not have sole use of bath/shower and toilet OR has no central heating.

Employment: Any member of the household aged 16-74 who is not a full-time student is either unemployed or permanently sick.

#### **4.3.4 Classifications available in the Controlled Access Microdata Files (CAMS)**

The CAMs data for England and Wales contains the 2004 Index of Deprivation scores and the values for the four domains that are consistent across all countries: Income; Employment; Health deprivation and disability; Education, skills and training.

For Scotland only the deciles into which the 2001 Index of Deprivations scores fall are available. The Index of Deprivation is based on super-output areas in all countries.

##### *Indices of Deprivation 2004*

[http://www.warrington.gov.uk/council/facts/Indices\\_of\\_Deprivation\\_2004.asp#1](http://www.warrington.gov.uk/council/facts/Indices_of_Deprivation_2004.asp#1)

<http://communities.gov.uk/>

##### *Scottish Index of Multiple Deprivations*

<http://www.scotland.gov.uk/stats/simd2004/>

<http://www.scotland.gov.uk/library5/government/glsimd-00.asp>

#### *4.4 Confidentiality and disclosure control*

Recent years have seen a sharp increase in concern over confidentiality – not just in the statistical offices of the UK but also in NSIs around the world. In part this is influenced by a recognition of the increasing amount of information available through stored databases and on the web and also the increasing power of search engines and data mining techniques. It also reflects increased concerns about privacy and the legal requirements of the census offices to ensure that the guarantee of confidentiality given on the census form is upheld. This increased concern over disclosure risk has resulted in a decision that small cells from the

tabular output from the 2001 Census for England, Wales and Northern Ireland should be adjusted to either zero or 3. It has also resulted in less detail being released in the 2001 SARs.

The ONS protocol for Data Access and Confidentiality sets out the overarching principles that relate to confidentiality. The general framework by which ONS protect census data is given at [www.statistics.gov.uk/census2001/discloseprotect.asp](http://www.statistics.gov.uk/census2001/discloseprotect.asp).

#### **4.4.1 Confidentiality and disclosure control in the 2001 SARs**

*Protecting confidentiality: statement on ONS web site*

The ONS web site [www.statistics.gov.uk/census2001/sar\\_update.asp](http://www.statistics.gov.uk/census2001/sar_update.asp) explains the confidentiality protection for the SARs as below. (this page was last revised on 30 September 2004).

"The Census Offices have a clear, well published, goal for protecting the confidentiality of individual information:

...In releasing statistics from the Census, all possible steps will be taken to prevent the inadvertent disclosure of information about identifiable individuals and households.

The Registrars General also have a legal obligation not to reveal information collected in confidence in the Census about individual people and households, and have given public assurances about what this means in practice. In presenting very detailed results from the Census, protecting individual information is of key importance. Traditionally the confidentiality of Census output is protected by a combination of disclosure control methods.

As well as the legal aspect of disclosure control ONS has also stated in the 2001 Census Disclosure Control advisory group paper AG0106 that:

"Maintaining the confidentiality of individual data underpins the trust that exists between data suppliers and any agency that acts as custodian of information about them. At ONS we are fortunate that businesses and the public have confidence that their information is securely held and that we do not release any data that could identify an individual. It is essential that this trust be maintained....".

Protecting the confidentiality of details about individual people becomes less simple with each Census, as the amount of accessible and publicly available information about individuals increases. More information can be matched statistically with the Census, and electoral rolls are more widely used in electronic form. Alongside this, for the 2001 Census a larger

range of small area statistics has been released, notably because some key measures which were previously obtained from 10 per cent samples were available in 2001 for the whole population. A much wider range of small area information is being published through Neighbourhood Statistics, from public records as well as the Census.

Since 1991 the internet has transformed the potential for making census results widely accessible to citizens. Changing attitudes to the trust in which public agencies are held and concerns about the importance of privacy of personal information also place new and more onerous demands on bodies responsible for protecting such information supplied in confidence.

The general strategy for ensuring the statistical confidentiality of 2001 Census output was stated in the Government's March 1999 White Paper *The 2001 Census of Population*:

"Precautions will be taken so that published tabulations and abstracts of statistical data do not reveal any information about identifiable individuals or households. Special precautions may apply particularly to statistical output for small areas. Measures to ensure disclosure control will include some, or all, of the following procedures:

- restricting the number of output categories into which a variable may be classified, such as aggregated age groups
- where the number of people or households in an area falls below a minimum threshold, the statistical output - except for basic headcounts - will be amalgamated with that for a sufficiently large enough neighbouring area and/or
- modifying the data before the statistics are released."

These considerations have led ONS to reassess how much detail be released from the 2001 Census. Additional measures have been introduced for tabular output and some restrictions in detail have been applied to the SARs.

#### *Disclosure Risk Assessment*

The Economic and Social Research Council, through the Cathie Marsh Centre for Census and Survey Research (CCSR), made a request for 2001 SARs. They also asked ONS to consider the following enhancements to the 1991 SARs specification:

- reduce the threshold for the Individual SAR from 120,000 to 90,000 population
- increase the sample size for the Individual SAR from 2 per cent to 3 per cent

- changes in detail given to some of the variables, for example ethnic group, family type and professional qualifications, to reflect changes in the information collected in 2001
- add extra variables (to reflect the new questions asked in the 2001 Census)

These proposals are based on the paper by Dale and Elliot; 'Proposals for the 2001 SARs: an assessment of disclosure risk'. This paper assessed the risk of disclosure from the SARs and concluded that the risk was very low. It suggested that the 1991 assessment of risk was pessimistic and there was scope for a decrease in the threshold and an increase in the sample size of the individual SAR.

ONS carried out further analysis to assess the risk. In particular, ONS recognised that a risk assessment for the country as a whole would not necessarily allow it to meet the commitments it has made to every individual who completed a Census form. In particular, some individuals are more easily recognisable in the population than others. The Census Offices have a responsibility to protect everyone's information, not just the majority.

ONS also considered how an attempt could be made to identify an individual. It considered what additional information and data would be available to users of the SARs (regardless of whether it was in the public domain) and whether this information could be used to identify an individual in the SARs.

The main elements of the analysis were:

- an analysis to determine whether or not a variable should be collapsed, similar to the analysis carried out in 1991. See The 1991 Census User's Guide, Chapter 5.4.4
- an analysis of the number and proportion of unique individuals in the sample who are also unique in the population. This looked at the total population as well as groups within it
- a qualitative assessment of the risk that an individual within the SARs can be identified by matching the SARs against an external dataset.

This analysis showed that grouping of age, ethnic group and occupation substantially reduced the risk of identifying an individual from the sample. It also showed that the sample size could be increased from 2 per cent to 3 per cent.

ONS also looked at the risk of identifying individuals by matching databases against other sources and whether or not some of the variables may be able to help in confirming the identity of individuals. Variables

such as the area classification, communal establishment type and family type were all found to increase the risk significantly by substantially narrowing down the location of an individual or groups of individuals in the population. These variables would either need to be excluded from the SARs or grouped into fewer bands.

It is likely that a small number of uniques will remain in the SARs sample once these checks have been completed. In order to further reduce the risk of identification of an individual ONS will carry out perturbation of the risky records using the PRAM technique (post-randomisation method). This will consist of changes to certain values in these records which may be applied by means of record swapping or imputation.”

The 2001 Samples of Anonymised Records has been subjected to a more extended analysis of disclosure control than that used with the 1991 SARs. A number of scenarios were set up as the likely routes through which an attack on the microdata might be made. For example, one scenario used the information that a journalist may be able to obtain on an individual whilst another related to the information held on databases used by commercial companies. Sets of key variables – which could be used to match to the same variables in the SARs - were identified, based on these scenarios. Then, using the full 2001 Census data, the records in the SAR that were population unique on the key variables were determined. The percentage of records that were population unique was used as the primary measure of risk.

Considering population uniques meant that only matches which were certain were counted as a risk. Matches which were correct by chance were not counted as a risk. After the proportion of population uniques had been reduced by collapsing variables, the most risky records were perturbed by changing the values of one or two variables. To control information loss the proportion of records perturbed was small. At least two-thirds of the population uniques in the file were perturbed. A more detailed discussion is available in PDF format.

#### **4.4.2 Perturbation in the SARs**

This section is based on a paper by Gross, Guiblin and Merrett presented at the Open Meeting on the 2001 SARs, September 2004. Details can be found at <http://www.ccsr.ac.uk/sars/guide/2001/pram.pdf>.

Recoding variables determined as risky is the main disclosure control method used for the SARs. However, there comes a point where further recoding will cause a large decrease in the information released for little decrease in disclosure risk. At this point ONS decided to use PRAM (Post-Randomisation Method) as a perturbative microdata disclosure control technique which can be applied to categorical variables. The values on

some categorical variables for certain records in the microdata file are changed to a different value according to a prescribed probability mechanism. Each new value may or may not be different from the original value. The key aspect of the PRAM method used is that the method conserves the original frequency distributions, while minimising the loss of information.

The number of records PRAMmed for each variable is given in Table 1. Values which were PRAMmed were flagged as imputed, but not so as to distinguish them from already imputed values.

**Table 1: Number of records PRAMmed for each variable by Country**

<b>Variable</b>	<b>England and Wales (1,626,324)</b>	<b>Northern Ireland (50,889)</b>	<b>Scotland (164,307)</b>
Age	7,510	302	961
Distance to work	2,414	134	Not PRAMmed
Marital Status	7,939	486	95
Number of cars in the household	2,589	128	57
Number of earners in the household	726	53	36
Number dependent children in hhold	2,157	111	Not PRAMmed
Number of residents in the household	2,554	114	40
Primary economic position	12,977	562	364
Tenure	4,348	245	96
Workplace	240	12	Not PRAMmed
Address last year	5,251	258	Not PRAMmed
Ethnicity	5,421	Not PRAMmed	271
Industry	7,143	237	147
Long term illness	313	Not PRAMmed	Not PRAMmed
Occupation	5,324	217	429

### **Effect of PRAM on data quality**

Three aspects of data quality were examined.

1 The invariance property – preservation of the univariate. This was checked by looking at univariate frequency tables pre- and post-pram and looking at the transitions (cross frequency between original variable and PRAMmed variable).

2 Preserving the multivariate frequencies within subgroups of variables is checked looking at multi-way tables. To achieve this criteria we used a series of control variables defining strata within which Pram was performed (e.g. Age within strata defined by workplace, Econprim and marital status).

3 Preserving the relationship between PRAMmed variables and non PRAMmed variables. This was not controlled. But the damage has been assessed by comparing frequency tables before and after PRAM.

The conclusions are:

1 The univariate distributions for PRAMmed variables were not damaged. The preservation of the frequencies worked pretty well. This means that the optimisation process worked well.

2 The multivariate distributions between variables involved in the PRAM process (PRAMmed and control variables) worked well too. This means than the stratification and the control on transition were efficient.

3 The assessment of the damage on distribution between variables involved and not involved in the PRAM process was measured by comparing tables before and after PRAM and by comparing the impact of PRAMming relative to the sampling error. The ratio between the relative error due to PRAM and relative sampling error was calculated for each cell. When the ratio is lower than 1 the additional error due to PRAM can be considered as acceptable. PRAM is more damaging (relative error due to PRAM > Relative Sampling Error) for cells which have low frequencies.

Below is an example of the loss of information affecting a three-way table: industry/sex/ethnic group. It showed a much larger loss of information than other tables examined. For each cell size of this 3-way table of 181 cells we measure the ratio between the error due to PRAM and the Sampling Error.

Cell Frequency Before PRAM	0-5	6-10	11-20	21-40	41-90	91-150	150-500	500+
Number of cells where Ratio >1	6	3	9	4	2	0	4	3

PRAM is more damaging for cells with low frequencies. Six cells with a frequency less than 5 have been more damaged by PRAM. (Note: an increase of 1 person after PRAM in a cell of 5 before PRAM represents a change of 20 per cent but an increase of 1 person after PRAM in a cell of 50 before PRAM represents a change of 2 per cent).

#### **4.4.3 Disclosure control report- Special Licence Household SAR (Source: ONS)**

##### *Introduction*

The original SARs project plan for the Household SAR was to produce a licensed file that would be accessible under the End User Licence at the UK Data Archives, the same way as for the Individual Licensed SAR and the Small Area Microdata (SAM) file. The original disclosure risk assessment of the Household SAR indicated that the file presented a very high risk of disclosure, with a high percentage of the households and the individuals within those households being population unique. Possible recodes were suggested to reduce the level of disclosure risk to a level acceptable for an End User Licence file. However implementing the recodes would render the file of little value to users, largely because age was the main variable to be recoded. To provide the user community with a file that met their requirements, single years of age for all household sizes, meant producing a more disclosive file which could not be released under the existing End User Licence arrangement.

At the time the Household SAR was being assessed, ONS was also reviewing its policy on microdata access generally and approved access to more detailed datasets for social surveys which allowed a combination of i) statistical disclosure control methods with ii) legally binding agreements, to protect data confidentiality. Approval to use this ONS Special Licence for the Household SAR was sought and approved. The use of the ONS Special Licences enables a more useful Household SAR to be produced. The overall protection of the data is thus partly through design, and partly through licensing, with control over the user and their intended use.

In the proposal to use the ONS Special Licence for the Household SAR, scenarios were outlined which the released data should be protected against. The consensus was that the risk of deliberate attempts at matching by researchers was very low. In other words the process of approval for the Special licence, the accountability of the institution and the severe impact of penalties for improper use allows us to trust the researcher, and provides protection against intentional confidentiality breaches. The physical security arrangements required by the licence are also considered adequate. The concern was the likelihood of careless or negligent actions by the researcher, particularly where the requirements

of physical security are contrary to their normal work habits. Failure through negligence to strictly adhere to the terms of the licence conditions could lead to unauthorised access and is the main difference in risk between the Special licence and access under a safe setting. It was also agreed that the data should be protected against spontaneous recognition of publicly well known individuals both by researchers and any unauthorised access by a third party.

### *Assessment of disclosure risk*

The assessment of the disclosure risk of the special licensed Household SAR has involved examining two strands, firstly assessing the risk from a private database match and secondly assessing the risk of spontaneous recognition of publicly well known individuals.

#### (i) Private Database

A risk analysis of the Household SARs was carried out to provide a quantitative measure of the disclosure risk. The risk measure used is the percentage of population uniques in the file at household and individual level. The characteristics used to define unique records are determined by key variables, which in turn depend on the intruder scenario being considered. Key variables are a set of variables that can be used to identify a particular unit (e.g. household or person) for a given intruder scenario. A unit record in a microdata file that is unique in the population on a set of key variables is highly likely to allow the unit corresponding to that record to be successfully identified, and all the information on the file about them made available to the intruder.

Households are defined as unique in the population if the particular combination of characteristics of the individuals in the household are not found in any other household. If a household unit is a population unique, all individuals within the household are also counted as population uniques, as they can be identified through the household. The chances of a household being population unique increase with the size of the household. Variables coded with more detailed categories will generally give rise to more population uniques than more aggregated coding.

The risk analysis uses the private database scenario to maintain consistency with the Individual SAR. The private database was slightly modified to adjust for the household structure. The modified key then consists of:

Household variables: household size, country, tenure, number of cars

Individual variables: age, sex, marital status, primary economic position  
Various coding schemes were considered for the Household SAR.

The coding scheme which has been adopted for the Household SAR is:

Age is in two year bands, and top-coded at 80 years

Marital status has 5 categories,

England and Wales are combined, and

Other key variables coded as for the SAM.

Remaining variables coded as for the Individual SAR

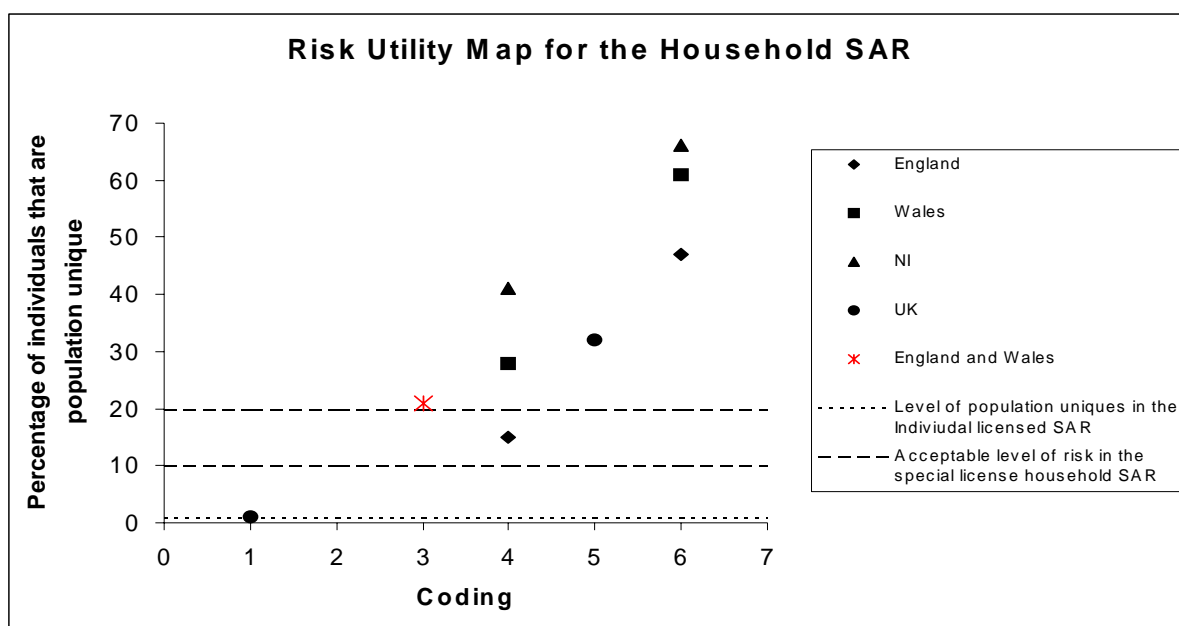
No data is provided for Scotland and Northern Ireland as the disclosure risk for these countries was deemed to be high even under special licence conditions.

The level of population uniques in the file was 12% of households and 21% of individuals. This compares to 1% of population uniques under the End User Licence. Table 1 shows the risk measures for all the coding options considered. These are converted to the risk-utility framework in Figure 1 (as used in Duncan et al, 2001). Coding schemes have been ordered by the amount of detail available, however as we have not used a quantitative measure, the data utility axis is not to scale. The map shows disclosure risk decreasing as data utility is reduced. The lowest risk file (at 1%) with 10 year age bands and households over size 9 removed did not meet the needs of data users.

Most households of size 6 and over were found to be population unique. Figure 2 shows the percentage of households which are population unique for 3 coding schemes by household size. For individual years of age and sex it can be seen that the percentage of households which are population unique substantially increases from size 4. For the modified private scenario the percentage of households which are population unique substantially increases at size 6. However, for the coarsest coding scheme the percentage of households which are population unique does not substantially increase until size 9.

The level of risk as shown by coding scheme 3 in Table 1 was felt to be acceptable given the greater reliance on contractual arrangements provided by the Special Licence and provided a good balance between disclosure risk and data utility.

Figure 1. Risk-Utility Map for the Household SAR



Note: Data Utility is expressed in the coding schemes as explained in Table 1

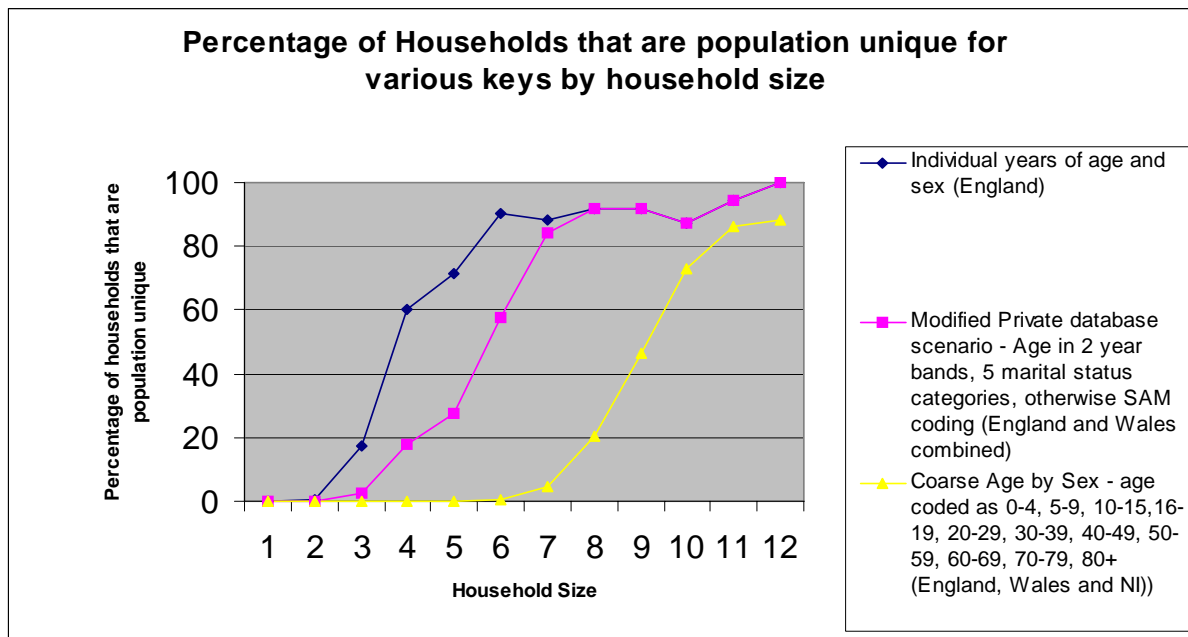
Table 1: Risk for the Household SAR with different coding options (private database scenario)

Country	Coding	Number of individuals	Number of households	Percentage of households that are unique	Percentage of individuals that are unique
UK (England, Wales and Northern Ireland)	Age in 10 year bands, Sex and Marital Status in 4 categories and up to size 9 households (1)	54266617	22328668	0.5	1
England and Wales	Modified private database scenario (SAM coding except age in two year band and Marital status in 5 categories) and England and Wales combined (3)	520636	216686	12	21
England	Single years of age and sex (4)	491500	204613	8	15
Wales	Single years of age and sex (4)	29136	12073	16	28
Northern Ireland	Single years of age and sex (4)	16895	6241	25	41
UK (England, Wales and Northern Ireland)	Modified private database scenario (SAM coding except age in single years) and countries combined (5)	528423	222927	19	32
England	Modified private database scenario (detailed coding) (6)	491500	204613	31	47
Wales	Modified private database scenario (detailed coding) (6)	29136	12073	45	61
Northern Ireland	Modified private database scenario (detailed coding) (6)	16895	6241	51	66

(1) these are population figures, but will provide a good estimate for the sample.

(4) considers only two variables, and is not a full scenario analysis

**Figure 2:** Percentage of households that are population unique for various keys by household size.



(ii) Spontaneous recognition of individuals who are well known publicly

This risk assessment looks at the risk of being able to identify well known individuals in the file. It was decided that it would be difficult to identify unusual households as few are well known by virtue of household characteristics. The analysis concentrated at looking for individuals who might be well known publicly at the national level.

Note that for the special licence we do not aim to protect the file against recognition of a "community of acquaintances", i.e. individuals or households that might be identified by particular "intruders". The researchers agree not to attempt to identify individuals, under the terms of the licence.

We were concerned only with those variables on the file that are visible and traceable. The 7 variables for the modified private database cross-match scenario are Household size, Tenure, Number of cars, Age, Sex, Marital Status and Primary Economic position. Of these we considered only household size, age, sex and Marital status would be used to try and identify a publicly well known individual. Other key variables that we felt might lead to recognition of a publicly well known individual on the Household SAR, which are not contained on the above private database scenario were Ethnicity, Religion, country of birth, occupation and International standard classification for occupation (ISCO). Thus, the variables that we considered for spontaneous recognition of publicly well known individuals were:

Household size - top coded at size 12  
Age - 2 year bands  
Sex - 2 categories  
Marital Status - 5 categories  
Ethnicity - 16 categories  
Religion- 9 categories  
Country of birth - 16 categories  
Occupation - SOC minor 81 categories  
ISCO - 3 digit level.

To examine the risk of being able to identify well known individuals two analyses were undertaken, these were:

- The univariate distributions of the visible and traceable variables not on the modified private database scenario were examined at the national level. We were looking for categories of variables with low numbers in particular, as this gave us an indication of where the disclosure risks were.
- Identify all population uniques in the sample for all three way combinations that include occupation. We considered that occupation was the main variable that in combination with other variables would contribute to recognising a well known individual at a national level. In our judgement it was sufficient to restrict the analysis to occupation and 2 other variables. We were not considering all possible combinations of the 9 visible and traceable variables.

The results of these analyses showed that it would be very difficult to identify a publicly well known individual based on the level of detail provided in the Household SAR. Therefore for variables not on the private database scenario it was decided they can be coded to the same level of detail as for the Individual SAR. There were however 6 records in the Household SAR that were population unique for a combination of an occupation typical of a well known individual and two other variables. The most identifying variable for an individual was changed manually to protect the 6 records.

#### *Disclosure control methods applied*

Recoding has been the main disclosure control method applied to the Household SAR. Only those variables which are on the private database have been subject to recoding and in addition no individual level information is provided for households of size 12 or more. In addition to the recoding a small amount of perturbation has been applied to the data to protect confidentiality, using the same methods as for the Individual SAR see Bycroft et al (2005).

## (i) Recoding

For the variables on the modified private database scenario some recoding was applied:

Tenure - this was recoded from 10 categories to 3.

Number of cars - this was recoded from 5 categories to 3.

Age - this has been banded into two year age groups and top-coded at 80.

Marital status - this was recoded from 6 categories to 5.

Primary Economic Position - this was recoded from 16 categories to 4.

Household Size - this was top coded at 12. For households of up to and including size 11, there will be one record for each member of the household. For households of size 12 or more only 1 record providing summary information will be provided for the household.

For full details of the coding of the variables please see appendix A.

### *Individual and household level edits*

After both the perturbation had been applied to the 5% of individuals in large households and the modifications to the 6 risky records from the visible and traceable analysis have been made, individual and household level edits were checked to ensure that no invalid combinations have occurred.

The individual level edits that were checked were the same as those agreed for the Licensed Individual SAR. These edits were derived based on the edits used by Census in creating the 2001 Census database and some additional edits that we felt should be checked. These edits checked to ensure that no invalid 'individuals' were created such as a 2 year old married person.

The household level edits that were checked were (i) Age difference between spouses should not be greater than 30 years, (ii) Age difference between Parent and child should be 16 years, and (iii) Age difference between Grandparent and Grandchildren should be 32 years or greater. These edits were used to ensure that the perturbation applied to the age did not create any new extreme households.

The results of running both the individual and household level edits were that only one new additional failed edit occurred through the perturbation. All the other failed edits already existed in the original data so it was decided that these would not be corrected.

The Edit list contained some edits that were failing on the original census records. We recommend that the edits be revised for the 2011 SARs. Better coordination at an earlier stage between the edits used for the Census and those for the SARs would assist in the development of the SARs.

### *Conclusions and Recommendations for 2011*

The Household SAR was the first ONS dataset to undergo a quantitative assessment of disclosure risk for access under the Special Licence. Decisions were made without the benefit of any previous experience of the new licence.

For 2011 there is a need to reassess the balance between the protection provided by Special Licence and the recoding of the data. If ONS is comfortable with the Special Licence arrangement and there is a demand from researchers for more detailed data then, for example it may be possible to provide single years of age.

#### **4.4.4 Disclosure control report- Small Area Microdata (SAM)**

(Source: ONS)

##### *Introduction*

The Small Area Microdata (SAM) file is a 5% sample of individuals from the 2001 Census. Geography area identification is at Local Authority level for England, Wales and Scotland and Parliamentary Constituencies for Northern Ireland. The case for the Small Area Microdata file was put to ONS in January 2001 by Tranmer et al (2005). Users of the 1991 Individual SAR had noted that the geographical units identifiable were often larger than desirable for the type of analysis that they wished to conduct.

The SAM will be accessible under conditions similar to those employed for the Individual Licensed SAR. As the SAM is similar to the Individual licensed SAR, except that it is larger and at a finer level of geography, and similar methodological practices could be used in the production of this file. This has enabled a smoother and quicker production process than for the other SAR products. As with the Individual Licensed SAR there has been a trade off between the constraints of confidentiality and the amount of individual level detail that could be provided in the SAM.

##### *Assessment of disclosure risk*

The assessment of the disclosure risk of the licensed SAM has involved examining two strands, firstly assessing the risk from a private database

match and secondly assessing the risk of spontaneous recognition of individuals.

A risk analysis of the SAM was carried out to provide a quantitative measure of the disclosure risk. The risk measure used was the percentage of individuals in the file which were population unique. The characteristics used to define unique records are determined by *key variables*, which in turn depend on the intruder scenario being considered. Key variables are a set of variables that can be used to identify a particular unit (e.g. household or person) for a given intruder scenario. A unit record in a microdata file that is unique in the population on a set of key variables is highly likely to allow the unit corresponding to that record to be successfully identified, and all the information on the file about them made available to the intruder.

#### (i) Private Database

The risk analysis uses the private database scenario to maintain consistency with the individual SAR. The private database contains the following variables: Local Authority District (Parliamentary constituencies for NI), Age, Sex, Distance to Work, Marital Status, Number of Cars, Number of Earners, Number of dependent Children, Number of Residents, Primary Economic position, Tenure, Workplace

Three coding schemes were initially considered for the SAM these being a fine, coarse and a coarse plus (coarse specification plus two additional recodes) specification which resulted in 26.5%, 3.5 %, 2.5% of population uniques for England and Wales. The coarse plus specification was decided on as the percentage of population uniques was similar to that for the individual SAR before perturbation was applied. The coarse plus specification was put out for consultation with users and the resulting specification for the SAM can be seen in appendix A. This specification resulted in 2.5%, 3.1% and 4.2% of records being population unique for England, Wales and Northern Ireland respectively.

As part of the disclosure risk assessment of the SAM a special uniques assessment was run. The special uniques assessment was used to decide which records and which variables would be selected for perturbation. For further details on the special uniques methodology see Elliot (2004).

The Data Intrusion Simulation metric (DIS) is a file level measure of risk that gives the probability that a match made by an intruder between an individual and a sample unique in the microdata file is correct. The latest special uniques algorithm combines the SUDA score and the DIS metric to generate estimated per record matching probabilities.

The first step of the special uniques assessment is a SUDA score for each record and this is based on the number and size of minimal sample uniques (MSU). A MSU is a set of variable values which is unique in the sample and for which no subset is unique. Each record may have a number of MSU's within a given key variable set. Using the information about the number and size of the MSU's a SUDA score for each record is derived. This score is then heuristically combined with the output from the DIS metric to give a per record matching probability (dis\_suda score).

For the SAM the special uniques assessment identified all the minimal sample uniques for that record within the private database cross-match scenario. As we had the population data it was possible for us to identify which records in the SAM were population unique for the private database scenario. We then used the results of this special uniques assessment to grade these population uniques, with the highest score being given to the "most" risky population uniques. A population unique resulting from a small number of variables will have a higher score than one that results from a larger number of variables. We use this ranking of records to select the highest risk records for perturbation.

The special uniques assessment also told us which *variables* were contributing the most to the risk for each record; this is based on the number of times a variable occurs in the set of minimal uniques for that record. We used this information in selecting which variable will be prammed for a record. We experienced some difficulty with running the special uniques analysis on such a large file with many variables including LA. We solved this problem by splitting the files by GOR and looking for an MSU of size 6 or less out of 12 variables in the scenario.

## (ii) Spontaneous Recognition

The SAM as well as being protected against a private database attack should also be protected against spontaneous recognition and by this we mean that an intruder would be able to recognise people they know personally or people who are in the public eye. Spontaneous recognition is seen to be unintentional disclosure and we have viewed it as requiring only a small amount of information to be known about an individual. For the purposes of this analysis we have defined the risk of spontaneous recognition as being population unique on 4 variables. Geographies GOR and LA (or Parliamentary Constituency), are one of the visible variables, so that high risk records will be unique at LA level plus three other variables, or at GOR level with four variables. We note that any record unique at GOR level on three variables will also be unique in their LA.

We were concerned only with those variables on the file that are visible and traceable. The private database scenario contains some of the visible and traceable variables but there are others in the SAM which are not

contained within this scenario. The variables that we considered to be visible and traceable on the file are:

### **Household level**

Local Authority (Parliamentary constituencies for NI)  
Accommodation type  
Family type  
Lowest floor level of household living  
Number of rooms  
Number of Cars  
Number of Earners  
Number of Dependent Children  
Number of Residents  
Density  
Accommodation self-contained.

### **Individual level**

Age  
Sex  
Marital Status  
Distance to work  
Primary economic position  
Workplace  
Country of birth  
Ethnic group  
Transport to work  
Migration indicator  
Religion  
NS- SEC  
Community background (Northern Ireland only)  
Professional qualification  
Supervisor/Foreman  
Limiting Long term illness  
Size of Work force

To investigate the risk of spontaneous recognition two analyses were undertaken these were:

- The univariate distributions of the visible and traceable variables not on the private database scenario key were examined by LA. We were looking for categories of variables with low numbers in particular, as this gave an indication of where the disclosure risks were.
- The special uniques analysis was run using a Minimal Sample Unique (MSU) of size 4 for the visible and traceable variables listed above. The MSU of 4 means that all possible combinations of size 4 out of the total 28 variables will contribute to the assessment of risk. The special uniques analysis produced a DIS-SUDA score for each sample

unique record. The DIS-SUDA score estimates the probability that a match against the record is a correct match. From the special uniques analysis we were able to tell which of the variables is contributing the most to the number of MSU's and which variable values were the most risky.

The results of examining the distributions of the visible and traceable variables by LAD showed that there were very few variables with small counts. The DIS -SUDA scores for the samples uniques were all low indicating a low chance of being population unique. These results were verified by taking the top 10 variables (LA, ethnicity, religion, age, NS-SEC, transport to work, migration indicator, country of birth, work place and distance to work) contributing most to the risk and examining all four way combinations of these variables in the population to see if the combination was population unique. We found that the probability that a sample unique is a population unique was small, confirming the low DIS-SUDA scores. Therefore based on this we concluded that the level of detail of the visible and traceable variables in the SAM was appropriate and no further disclosure control for these variables was required.

### (iii) Risk Assessment for Scotland

Scotland has a different disclosure risk problem as it is possible to use tables which are in the public domain to confirm whether a record in the microdata sample is population unique. This is due to the fact that Scotland chose not to apply small cell adjustment to their tabular outputs which means that some values of 1 are present in published Scottish tables. Scotland's disclosure risk assessment has consisted of producing three way population tables for all variables on the SAM specification and then checking whether any population uniques in these tables occurred in the sample. This risk assessment was conducted by Sam Smith at the University of Manchester.

#### *Disclosure control methods applied*

Recoding has been the main disclosure control method to be applied to the SAM. However there came a point when further recoding would have resulted in little reduction in disclosure risk but large information loss. For the remaining high risk records in the SAM some perturbation was applied.

### (i) Recoding

Due to the smaller geography on the SAM compared to the Individual SAR, it was necessary to substantially reduce the individual and household level information. Some information was completely removed from the file such as the Standard Occupation Classification of the

individual and the industry that they worked in. For full details of the final coding used in the SAM please see appendix A.

## (ii) Perturbation

The results of the special uniques analysis were used to efficiently target the perturbation to the highest risk records and highest risk variables. The special uniques analysis ranks sample uniques in the file by what is called a DIS/SUDA score. The methodology used to perturb the records in the SAM is based on the Post Randomisation Methodology. For further information on the PRAM methodology used please see Bycroft and Merrett 2005. PRAM was applied only to records that exceeded a threshold for the DIS/SUDA score and were population unique for the private database scenario.

Records in the SAM which were imputed were flagged. We used the same flag to indicate whether a record had been subject to PRAM. This informs the user that the value is not obtained directly from a true response, but does not allow them to distinguish between the two processes. Therefore if an intruder comes across a flagged record they do not know whether it is a true value, perturbed or imputed. Northern Ireland and Scotland choose to remove the imputation and PRAM flags from their files before release to provide some additional protection.

## *Edits*

After the perturbation had been applied to the file, edits were checked to ensure that no invalid combinations have occurred. The edits that were checked were the same as those agreed for the Licensed Individual SAR . These edits were derived based based on the edits used by Census in creating the 2001 Census database and some additional edits that we felt should be checked. These edits checked to ensure that no invalid 'individuals' were created such as a 2 year old married person.

The results of running the edits were that no new additional failed edits occurred due to the perturbation. All the other failed edits already existed in the original data so it was decided that these would not be corrected.

## *Conclusion*

In conclusion the SAM is a valuable file providing detailed geography. The trade off for this detailed geography has been in the reduction in the detail of information provided for individuals. However the level of detail of the variables on the file is such that the file will still be useful for research purposes.

## *4.5 The accuracy and quality of data in the 2001 Census*

### **4.5.1 Sources of information**

Most of the information in this section is taken from the ONS web site at [www.statistics.gov.uk/census2001/methodology.asp](http://www.statistics.gov.uk/census2001/methodology.asp). Fuller information can be found on that site.

More information for Scotland can be found in 'Scotland: Taking Scotland's 2001 Census' at [www.gro-scotland.gov.uk/grosweb/grosweb.nsf/pages/cencr102](http://www.gro-scotland.gov.uk/grosweb/grosweb.nsf/pages/cencr102).

All this material remains Crown Copyright.

Census 2001 General report for England and Wales:

This reviews the entire Census operation from the early consultation and planning stages, to the production and dissemination of outputs and evaluation. It provides a wealth of detail about how the Census was carried out and what lessons have been learned to take forward in the plans for any future censuses. It is aimed at both the experienced and occasional user of census data, but it is hoped the wider public may also find it useful and informative. For the full report please see [www.statistics.gov.uk/StatBase/Product.asp?vlnk=14213](http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=14213)

The Census 2001 Quality report for England and Wales provides information about all aspects of quality relating to the 2001 Census. It provides an overview of the quality issues and the studies and analyses that have been carried out to improve the quality of Census data. The report deals with the life cycle of the Census project stage by stage, and then provides measures of each of the attributes of quality as defined by the European Statistical System. The final part describes the components of quality of the data for each Census question. In conjunction with the Census 2001 General report for England and Wales, it provides a comprehensive evaluation of the strengths and weaknesses of the Census operation. For the full report please see [www.statistics.gov.uk/downloads/census2001/census\\_2001\\_quality\\_report.pdf](http://www.statistics.gov.uk/downloads/census2001/census_2001_quality_report.pdf)

### **5.2 Measures of quality in the 2001 Census**

This section is extracted from ONS Census 2001 quality report for England and Wales. For more information see [www.statistics.gov.uk/downloads/census2001/census\\_2001\\_quality\\_report.pdf](http://www.statistics.gov.uk/downloads/census2001/census_2001_quality_report.pdf)

This section discusses the concept of quality that is often described as 'fitness for purpose' in terms of user needs. The Office for National Statistics (ONS) has adopted the European Statistical System for the description of quality, which is based on these attributes:

- Relevance
- Accuracy
- Timeliness
- Accessibility and clarity
- Comparability
- Coherence

The section defines each of these attributes and explains how they impact on Census quality. The aim is to combine the various processes undertaken during the Census that contribute to that quality attribute.

*Relevance:*

Relevance reflects the degree to which statistical information meets the needs and priorities expressed by users. Users need input into the topics, concepts and definitions underlying data to ensure its relevance. ONS has maintained close links with users, and undertook extensive user consultations with the identified communities including Central Government, Local Authorities (LAs), the health service, business and academics to gain an understanding of users' requirements for census information.

*Accuracy:*

Accuracy is the closeness between an estimated value and the unknown true population value. There is no single aggregate or overall measure of accuracy. However, it can be measured or described in terms of error, or the potential significance of error, introduced through sources such as coverage, response and processing.

The main errors that impacted on data accuracy were sampling and non-sampling error. As the 2001 Census was a measure of the whole population, there was no sampling error directly associated with it. However, as some under-enumeration occurred, sampling error was introduced by way of the imputation of additional people by the ONC process.

*Timeliness:*

Timeliness is the length of time between the date of the Census and the availability of data. The time lag between fieldwork and results should be minimised to permit the information to be of greatest value.

The 2001 Census was held on 29 April 2001 and the first results were published 17 months later on 30 September 2002. Data releases were

staggered so that headline results were released as quickly as possible, with more detailed statistics released later. This allowed time for additional processing where required.

*Accessibility and clarity:*

Accessibility reflects the availability of information, taking into account the suitability of the form the information is available in, the media of dissemination and the availability of metadata. The affordability of that information to users in relation to its value to them is also important.

Access to the main 2001 Census results is free and use is unrestricted. This is a major change from previous censuses, and reflects wider policies on access to government information.

*Comparability:*

Statistics are most useful when they enable comparisons across space and time. The need for comparability with the 1991 Census was a key factor in the design of the 2001 Census. However, changes in questions, concepts and definitions between 1991 and 2001 were necessary to take into account the need for harmonisation with other government surveys, to reflect changing customer requirements and to take account of new and improved data collection and processing methodologies.

Key changes between the 1991 and the 2001 Census included:

- Changes in population definitions, eg the enumeration of students at their term-time residence rather than their vacation address as in 1991.
- Changes in geographic boundaries between 1991 and 2001.
- Changes in the methodologies used. The 2001 Census was the first to adjust for under enumeration using the One Number Census process.
- Changes to the questions asked in the Census.

*Coherence:*

Coherence of data and information reflects the degree to which data can be logically connected across other data sets. Statistics are coherent if they are based on common definitions, classifications and methodological standards. The messages that statistics convey to users will clearly relate to each other and not contradict if the data is coherent.

ONS has been developing a programme of work to join up different statistics, ensuring a coherent, integrated presentation of data to users. Definitions have been harmonised across surveys, but there are differences in approach between, for example, self-completion questionnaires and interview surveys. Thus the Census questions to establish economic activity rates were somewhat more limited than those asked on the Labour Force Survey, where the interviewer can probe more deeply if required.

### 4.5.3 Response to the 2001 census

The proportion of people returning a census form in England and Wales was 94 per cent. In Northern Ireland it was estimated that 95.2 per cent of the population responded to the census and 95.3 per cent of the population in private households.

The total overall response for England and Wales was 98 per cent – including 4 per cent of the population estimated to be resident in households identified by enumerators but who were imputed. Table 1 shows the components of response at the 1991 and 2001 censuses.

<b>Table 1: Components of UK Census response and coverage rates for 1991 and 2001 - England and Wales</b>						
	<b>England</b>		<b>Wales</b>		<b>England &amp; Wales</b>	
	<b>1991</b>	<b>2001</b>	<b>1991</b>	<b>2001</b>	<b>1991</b>	<b>2001</b>
<b>A</b> <b>People on returned forms:</b> <b>Census Response Rate</b>	96	94	97	94	96	94
<b>B</b> <b>Other people in identified households</b>	2	4	1	4	2	4
<b>A+B</b> <b>Total overall response</b>	98	98	98	98	98	98
<b>C</b> <b>People not included on returned forms and people in wholly missed households</b>	2	2	2	2	2	2
<b>Total</b>	100	100	100	100	100	100
<b>Proportion of population covered in census results:</b> <b>Census Coverage Rate</b> <b>1991: A+B</b>	98	100	98	100	98	100

<b>2001: A+B+C</b>						
--------------------	--	--	--	--	--	--

**Note: The 1991 rates shown are subject to slight change, but this does not affect the conclusions to be drawn from this analysis.**

### *Response by age and sex*

Under-enumeration does not occur uniformly across all age-sex groups. Response rates were lowest for persons in their twenties, particularly men. Response levels by age-sex group for England and Wales as a whole varied from 98 per cent for females aged 75-79 to 87 per cent for males aged 20-24. A spreadsheet showing detailed rates for 1991 and 2001 is available from the ONS web site.

Census response has declined between 1991 and 2001 for most age-sex groups. Response rates have also declined for large-scale Government Surveys during the 1990s and response rates for censuses conducted in other countries have also fallen over the past decade, for example the level of under-enumeration observed in the censuses of both Australia and New Zealand rose between 1996 and 2001.

A summary of response patterns is given below:

- census response is lowest for the 20-24 age group for both men and women - for men this was also the case in 1991 but for women in 1991 the lowest census response was estimated to be for those aged 85+
- the largest difference in response rates is for Females 20-24 where response is 6 per cent lower in 2001 than 1991
- the response rate for children is much lower in 2001 than in 1991
- response rates were higher in 1991 than 2001 for both men and women across all age groups except for those aged 80 and over
- in both 1991 and 2001 there was a significant drop in response rates for those aged between 20 and 30
- generally there is a higher response rate for women than men, although this difference is smaller in 2001 than in 1991.

### *Response by area*

Under-enumeration also varied by area with lowest response rates for inner city areas where characteristics known to be related to census non-response are most prevalent - multi-occupancy and higher proportions of non-English speaking population etc.

Table 2 shows response rates for area types. The 2001 census response was lower in all area categories than 1991, with broadly similar proportionate drops across all areas with the exception of Inner London and Outer London, which have higher decreases in response rates. Inner

London had the lowest response rate in 1991 and recorded the largest absolute drop in 2001.

<b>Table 2 Census Response by area - England &amp; Wales</b>						
	<b>All people</b>		<b>Male</b>		<b>Female</b>	
	<b>1991</b>	<b>2001</b>	<b>1991</b>	<b>2001</b>	<b>1991</b>	<b>2001</b>
<b>Inner London</b>	88%	78%	86%	77%	90%	79%
<b>Outer London</b>	96%	90%	95%	89%	96%	90%
<b>Main Metropolitan areas</b>	94%	92%	92%	91%	96%	92%
<b>Other metropolitan areas</b>	97%	95%	96%	95%	98%	96%
<b>Non-metropolitan cities</b>	95%	94%	93%	93%	96%	94%
<b>Other non-metropolitan areas</b>	97%	96%	97%	96%	98%	97%
<b>Cardiff, Newport &amp; Swansea</b>	95%	93%	94%	93%	97%	94%
<b>Other Welsh areas</b>	97%	94%	97%	94%	98%	95%
<b>Total</b>	96%	94%	95%	93%	97%	94%
<b>Note: The 1991 rates shown are subject to slight change, but this does not affect the conclusions to be drawn from this analysis.</b>						

#### 4.5.4 The One Number Census

The 2001 Census aimed to maximise coverage and to make an accurate estimate of the people missed. The 1991 Census was thought to have had a substantially larger under-count than in previous censuses with about 2 per cent of the population of GB missed entirely and a further 1.6 per cent for whom records were imputed.

The One Number Census was designed to produce figures from the 2001 Census that are adjusted for under-enumeration and which are consistent across all forms of output and at the smallest geographical area. The term 'One Number Census' indicates a departure from the 1991 Census where preliminary figures from the census count were published and then later figures, adjusted for under-enumeration, were published. The One Number Census approach makes all adjustments as part of the census processing. Thus the One Number Census results in a database of the complete population for the UK from which all census outputs – including the SARs - are drawn.

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.

### *Step by step guide to the ONC*

The One Number Census process involved a number of stages:

- a Census Coverage Survey was designed and conducted independently of the census during May/June 2001
- records from the CCS were matched to those from the 2001 census
- populations of the sample areas were estimated from the results of the matching using dual system estimation techniques which enabled an estimate of those persons missed by both the census and the CCS to be made
- populations for each local authority by age and sex were then estimated using a combination of standard regression and small area estimation techniques
- households and persons estimated to have been missed by the census were then imputed to produce a fully adjusted census database and finally
- all population estimates were carefully quality assured using demographic analysis and comparison with aggregate level administrative data.

One of the key elements of the ONC was an independent follow-up survey, the Census Coverage Survey (CCS). This involved face to face interviews with a sample of 320,000 households from every local authority in the UK. But by combining the results of the census and the CCS, it was possible in 2001 to estimate the total resident population - the 'one number' - to a high level of precision, plus or minus 0.2 per cent.

### *The Census Coverage Survey (CCS)*

The CCS was specifically designed to enable census population counts to be adjusted for under-enumeration at the national, local and small area level. It consisted of a completely independent and intensive face-to-face survey of a sample of over 16,000 postcodes containing 320,000 households drawn from all local authorities in England and Wales. The sample design took into account the uneven distribution of under-enumeration across the country by stratifying by a 'Hard to Count' index based upon characteristics likely to be associated with under-enumeration, such as the number of multi-occupied addresses.

The CCS was operationally independent from the census enumeration exercise. The CCS sample postcodes were kept confidential, CCS

interviewers did not have any sight of the address lists produced in carrying out the census, nor the census forms returned in the area in which they were interviewing. The interviewers focused on making as many calls as necessary to achieve an interview and the timing of these calls was varied to maximise the probability of making contact.

The CCS in England and Wales achieved a response from 91 per cent of the households identified by interviewers. This is a high response rate for such a large-scale voluntary survey when compared to other national surveys. The survey succeeded in meeting its objective of identifying households and persons that had been missed by the 2001 census.

### *Quality assurance*

All the ONC population estimates were subject to rigorous quality assurance. The population of each local authority by age and sex were considered in a consistent and detailed manner - this involved comparison against diagnostic ranges derived from rolled-forward population estimates and aggregated administrative sources (such as Birth Registration and Pensions data). Where the ONC estimates fell outside of the diagnostic ranges, extensive checks of the ONC results were undertaken with respect to sample sizes, outliers etc and contingency action was taken if any issues were identified.

The quality assurance process included analysis for each local authority of a number of specific population subgroups known from 1991 to be prone to under-enumeration. These were full-time students, home armed forces, foreign armed forces and their dependents and prisoners. The estimates for these subgroups were compared with data from other official sources to determine whether the results were plausible.

### *Dependence between census and CCS*

For the ONC process to produce unbiased estimates of the population it is necessary for the census and Census Coverage Survey to be as independent of each other as possible. Practical arrangements were put in place to achieve this with census and CCS operations being kept entirely separate on the ground. If the two attempts at enumerating the same population are independent, it is possible to not only estimate those missed by either the census or CCS but to also estimate those missed by both - the dual system approach.

Through this approach, independence of process was achieved. However, there is an additional component of dependence which needs to be taken into account. This is dependence caused by the fact that those people who are difficult to count in a census are also difficult to count in a post-enumeration survey such as the CCS. This was expected and a methodology was developed to identify those areas where dependency

was marked and to adjust for that dependence. This added an additional 230,000 to the ONC population estimates for England and Wales as a whole.

### *Overcount*

Part of the CCS interview was also aimed at identifying any potential overcount in the 2001 census, that is persons incorrectly enumerated as resident at more than one address. Examples include second homes and children from broken homes living a proportion of time with each parent. Analysis of responses to the CCS indicated that the level of overcount in the 2001 census was negligible - less than 0.1 per cent of the population were estimated to have been counted twice.

The One Number Census Guide provides full details of how the ONC was conducted and is available at [www.statistics.gov.uk/census2001/onc.asp](http://www.statistics.gov.uk/census2001/onc.asp).

#### **4.5.4 Quality of responses**

In the 2001 Census a person was taken to exist if at least two of the name, date of birth and sex fields were completed. Generally, forms were accepted that contained a minimum of four items of information: name, date of birth, sex and marital status. Table 3, extracted from information on the ONS web site at [www.statistics.gov.uk/census2001/proj\\_qr.asp](http://www.statistics.gov.uk/census2001/proj_qr.asp), shows item non-response rates to topic areas for England and Wales.

It is evident that item non-response is lowest for age, sex and marital status and highest for company size and professional qualifications.

**Table 3: Item non-response rates, 2001 Census: England and Wales**

<b>Topic</b>	<b>England and Wales</b>
Age	0.5%
Sex	0.4%
Marital status	0.8%
Student flag	1.3%
CoB	2.5%
Ethnic	2.9%
Welsh Language	5.5%
Religion	7.6%
Health	3.1%
Carer	6.1%
Long-term illness	3.9%
Address 1 year ago	4.5%
Quals	6.2%

Prof quals	17.2%
Work last week	2.1%
Employment status	6.6%
Company size	13.9%
Occupation *	3.2%
Supervisor	6.8%
Industry *	7.8%
Workplace postcode	7.8%
Method of Travel	6.3%
Hours worked	8.0%
Relationship to Person 1	3.5%
Accommodation type	3.0%
Self-contained	3.9%
Rooms	5.4%
Bath/shower	2.5%
Lowest floor level	4.0%
Central Heating	2.2%
No. of cars	2.7%
Tenure	3.4%
Landlord	2.9%

In 1991 the Census validation Survey provided evidence of the accuracy – or at least the consistency – with which census questions were answered. The Census Coverage Survey did not fulfil this function and the only information available comes from the Census Test that was conducted in 1997. More details can be found at [www.statistics.gov.uk/census2001/proj\\_qr.asp](http://www.statistics.gov.uk/census2001/proj_qr.asp).

### *Edit and imputation*

As part of the planning to take account of under-enumeration through the One Number Census, much more information was imputed in the 2001 Census than for 1991. It is of great importance to analysts, particularly those using microdata, to understand the methods used for editing and imputation. For the SARs, all imputed records have been flagged to that users can choose whether or not to include them in any analysis.

The material below is drawn from [www.statistics.gov.uk/census2001/editimputevrep.asp](http://www.statistics.gov.uk/census2001/editimputevrep.asp). More details are available at that site.

An Edit and Donor Imputation System (EDIS) was devised for the 2001 census. Values would be set to missing for imputation if edit could not resolve an inconsistency. A person was taken to exist if at least two of the name, date of birth and sex fields were completed.

For the 2001 Census edit and imputation followed the principles below:

- all changes that were made would improve the quality of the data
- the number of changes to inconsistent data would be kept to a minimum
- as far as possible missing data would be imputed for all variables, so as to provide a complete and consistent database
- the system had to be relatively easy to develop and capable of processing large amounts of data automatically within short timescales.

## **Methodology**

EDIS can be sub-divided into five elements.

*Multi-tick rules* dealt with cases where more than one box was ticked but only one option was allowed. In some cases there was a rule for selecting one tick. If more than half the boxes were ticked or a set of priorities for accepting one tick could not sensibly be made, the answer was treated as missing and a value was supplied at the imputation stage.

*Range checks* were applied to prevent answers being outside an acceptable range. These were set to missing for subsequent imputation. Examples were households with 0 or more than 99 rooms, or with more than 20 cars, people with a date of birth before 1891 or after Census Day, who last worked before 1941 or who worked more than 99 hours per week.

*Filter rules* were applied to resolve some inconsistencies and to decide which fields should be set to 'No Code Required' where questions were answered but should not have been. For example, people under 16 or over 75 were not required to answer any of the employment questions. The variable Activity Last Week was also derived at this stage.

A set of Edit rules was applied to missing items or responses which appeared to be in error or inconsistent when compared with other data (such as married couples of the same sex, a child less than 13 years younger than its parents, or a married person under 16). These are known as hard checks.

In determining how to resolve such inconsistencies, the Fellegi/Holt principle of making the minimum number of changes was followed as far as possible. Thus if a person was under 16, married and had answered employment questions such as occupation, Age would be set to missing,

since the inconsistency could be resolved with the least change by imputing a value for Age between 16 and 74.

Edit also identified unlikely, but not impossible responses. In some cases rules were applied to eliminate these, for example, a purpose-built flat was considered unlikely to have more than 10 rooms, and for reasons explained below the value was set to 'Missing' for imputation. In others no further action was taken, e.g. where people under 35 were retired from paid work. The number of these 'soft checks' was reported but the data were not changed as a result.

All items which were missing after the Edit stage were dealt with by the Imputation component, which is described below.

*Imputation* was applied when there was no answer on the Census form, it failed the multi-tick rules or was invalid, or the filter rules or Edit marked it for imputation to resolve an inconsistency.

The principle of a Donor Imputation System is to search for a single donor household to supply all the missing variables in a recipient household. Exceptions are imputation for postcode of usual address one year ago and of workplace, which were carried out at a later stage than imputation for other variables.

The search looked at all records in an Estimation Area, a group of contiguous Local Authority Districts of about 500,000 population. The method searched for a donor using up to five matching variables, which were determined by the fields requiring imputation on the recipient record. Values were copied over from the donor household to fill the missing values on the recipient record. Consistency checks were then applied and the donor rejected if any check failed.

Potential donor households were scored using a second set of matching variables relating to all people in the household. In addition, potential donors were penalised if they had been used before as a donor or if any of their fields had been edited or imputed. A record could not be used as a donor if any of the fields to be imputed were also missing on the donor. If potential donors still scored equally, the donor geographically closest to the recipient was chosen. However, to improve efficiency of the searching procedure, if a suitable donor was found who lived within 5,000 metres of the recipient, this person was accepted and no further search took place to find a closer donor.

The intention was to use joint imputation where possible, i.e. selecting a single donor household to impute for all the people with missing values in a recipient household so as to preserve the joint distributions between variables. If a suitable donor household could not be found for joint

imputation, separate donors were sought to provide values for each person in the household requiring imputation, if necessary reducing the number of matching variables.

A fallback stage was also required as donor imputation failed to work for a few people. Most of these were imputed by testing possible values at random until one could be found which met the consistency criteria (a 'cold deck' approach). A small number of households could still not be completely resolved because of inconsistencies in age and relationships between people. As a final stage, if all else failed ('son of fallback') those containing up to eight people were completely replaced by synthetic households drawn at random from a set of the same household size, and households of nine or more people were corrected clerically.

The aim of imputation was to estimate the distribution of missing values accurately, so as to take account of any differences between the characteristics of respondents and non-respondents (non-response bias). It was not expected that the imputed values for every individual would be precisely accurate.

In comparison with 1991, EDIS was more comprehensive. It was applied to all variables, including qualifications, relationships, occupation, industry, hours worked, workplace address and means of transport to work, which were only analysed for a 10 per cent sample of households and communals in 1991.

There was some manual intervention in the 1991 processing system, such as clerical checking of missing or inconsistent items which exceeded certain tolerances. EDIS was almost entirely automatic as clerical intervention was limited to households of more than eight people which failed the fallback stage.

## **Edit**

A total of 13.7 million edits were carried out on the data for 11.8m people. The base population for EDIS was 49.4m people in England and Wales, including some 0.6m students living away from home during term-time for whom only a few demographic and relationship questions applied at their home address. The eight most frequently executed edits accounted for 91 per cent of the total. These were:

- |       |   |
|-------|---|
| 4.50m | Professional qualifications set to None where missing but educational qualifications was answered |
| 2.29m | Carer set to No where missing unless Activity Last Week was also missing                          |
| 1.66m | Workplace size set to 1-9 where person was self-employed  |

1.08m	Travel to work set to "work mainly at/from home" where workplace address was "mainly work at/from home"
1.03m	Supervisor set to No if missing, unless occupation was also missing
1.01m	Health set to Good if missing, unless Activity Last Week was also missing
0.59m	Professional qualifications set to missing if answered but educational qualifications was missing
0.40m	Missing Country of birth set to that of either siblings, parents or other related people in the household who have the same Country of birth

## Imputation

One or more items needed to be imputed for 13.8m people - that is 28.0 per cent of the population who returned Census forms. Of these, 4.7m were dealt with by joint imputation. 10.0m were imputed using individual imputation, including all those in single person households. 9.8m of the individual imputed cases used a donor household of the same size as the recipient's and the remaining 0.2m a household of different size. 0.4m people required imputation using the cold deck fallback method. Over 1m people had some items imputed by one method and some by another, hence there is some double-counting.

23.4 per cent of the population were used once as donors, 2.1 per cent twice and 0.1 per cent three or more times. In the SARs the variable EDISDONO provides information for whether an individual was used as a donor and, if so, how many times.

For household variables, 2.5m needed imputation, 11 per cent of all households. 0.08m were dealt with by fallback and the remainder by joint imputation. Almost all the donor households for joint imputation were used once each.

## Person variables

	Total (including imputed)	Non- response	Imputed	Non- response	Imputed
	000s	000s	000s	%	%
Age	49,359	262	278	0.53	0.56
Sex	49,359	199	219	0.40	0.44
Marital status	49,359	372	158	0.76	0.32

Student flag	49,359	622	641	1.26	1.30
Country of birth	48,848	1,211	829	2.48	1.70
Ethnic group	48,848	1,405	1,421	2.88	2.91
Welsh language	2,754	153	153	5.54	5.57
Religion	48,848	3,721	-	7.62	-
Health	48,848	1,525	531	3.12	1.09
Carer	48,848	2,967	693	6.07	1.42
Long-term illness	48,848	1,899	1,915	3.89	3.92
Address one year ago	48,848	2,198	2,213	4.50	4.53
Educational qualifications	35,367	2,187	-	6.18	-
Professional qualifications	35,367	6,094	-	17.23	-
Highest qualification	35,367	-	2,150	-	6.09
Working last week	35,367	737	-	2.08	-
Activity last week	35,367	-	1,301	-	3.69
Employment status	33,686	2,205	2,058	6.55	6.14
Workplace size	33,686	4,689	3,067	13.92	9.15
Supervisor	33,686	2,294	1,119	6.81	3.34
Occupation - currently working	21,741	694	759	3.19	3.48
Occupation - all ever worked	29,335	4,051	4,051	13.81	13.81
Industry - currently working	21,741	1,702	1,777	7.83	8.15
Industry - all	29,335	5,400	5,400	18.41	18.41

ever worked					
Workplace address	22,396	1,744	1,426	7.79	6.42
Method of travel	22,533	1,410	1,127	6.26	5.07
Hours worked	22,533	1,804	1,506	8.00	6.77
Relationship to Person 1	28,065	971	1,326	3.46	4.73

Note: The 'Total' column refers to the number of people in scope for the question, i.e.:

- Age, Sex, Marital status, Student flag: All people plus students who were counted at both their home address and term-time address in England and Wales
- Country of birth, Ethnic group, Religion, Health, Carer, Long-term illness, Address one year ago: All people (students counted at term-time address only)
- Welsh language: All people living in Wales
- Qualifications, Working last week: All people aged between 16 and 74
- Employment status, Company size, Supervisor: All people aged between 16 and 74 who had ever worked
- Workplace address, Method of travel, Hours worked: All people aged between 16 and 74 who were working in the week before Census day
- Relationship to Person 1: All people in households plus students also counted at home address less those who were entered as Person 1 on census form.

### *Age*

Age was not reported or was out of range (born after Census day or more than 110 years old) for 240,000 people. It was set to missing for a further 23,000 on grounds of inconsistency, mainly because people who were not single and who had answered three or more employment questions had their age captured as under 16.

The distribution of imputed ages followed that of the remainder of the population except for a shortfall among the 0, 6-15 and 76-80 age groups. This is primarily because some people were imputed as aged between 16 and 74 who may have been outside this age range because some employment questions had been answered. The shortfall in babies under 1 year old occurred where their address one year ago had not been stated as 'no usual address'. The effect in an area of 100,000 population

would typically be that two or three under 1's would have been imputed as over one.

### *Sex*

Sex was missing for 185,000 people and multi-ticked for 14,000, 0.4 per cent of the population in total. There were no edit actions which directly affected this question: if a husband and wife, or the parents of a child, were of the same sex the relevant relationships were imputed. A further 20,000 had values imputed by 'son of fallback'.

The sexes were imputed in the ratio of 51:49 in favour of females, very similar to the proportions among the remainder of the population. The accuracy of imputations was assessed by comparing the imputed values with people's names in a sample of areas. This showed that 75 per cent of imputations were correct. Among the incorrect values there was a very slight bias towards imputing females. The net effect would be to count four people out of every 100,000 as female rather than male.

### *Marital Status*

There were 373,000 missing or multi-ticked cases for marital status, representing 0.8 per cent of the population. 232,000 of these were children under 16 who were set to Single in edit. A further 6,000 under 16s had marital status changed to Single. Imputation was applied to the remainder. Married and Re-married were less likely to be imputed than among the remainder of the population.

### *Student*

Question 5 on the person schedule asked whether a person was a schoolchild or student in full-time education. 1.3 per cent of people failed to answer or multi-ticked the question, of whom 13 per cent were imputed as students compared with 21 per cent in the remainder of the population.

### *Country of Birth*

Country of birth was omitted by 2.5 per cent of people. Of these, 88 per cent were imputed as born in the United Kingdom, compared to 92 per cent in the remainder of the population. People born in Africa, Asia and North America were imputed in higher proportion than the remainder of the population.

### *Ethnic Group*

The non-response rate for ethnic group was 2.9 per cent. 89 per cent of these were imputed as White compared with 92 per cent in the remainder

of the population. There were higher proportions of imputed people in the Mixed, Asian and Black groups.

	Imputed		Total (including imputed)	
	000s	%	000s	%
White	1,260	88.7	45,065	92.3
Mixed	24	1.7	605	1.2
Asian	80	5.6	1,925	3.9
Black	43	3.0	868	1.8
Chinese and other	13	0.9	382	0.8

### *Welsh Language*

The question asking whether people could understand spoken Welsh, or speak, read or write the language, was asked of all people living in Wales. There was a 5.5 per cent non-response rate. No knowledge of Welsh was imputed slightly more often than for the remainder of the population.

### *Religion*

As the question on religion was voluntary, non-responses were not imputed but will appear in tables as 'not stated'. The national non-response rate was 7.6 per cent.

### *General Health*

This question asked whether over the last twelve months a person's health had on the whole been good, fairly good or not good. The non-response rate was 3.1 per cent, but an edit rule set the value to good unless Activity Last Week was also missing. This reduced the number requiring imputation to 1.1 per cent. Among these people, Fairly Good and Not Good were imputed slightly more frequently than in the remainder of the population.

### *Carer*

Question 12 referred to voluntary help or support given to family members, friends or neighbours. The rate of non-response was 6.1 per cent. Missing values were set to No by an edit rule unless Activity Last Week was also missing, and children under 5 were also assumed to not be providing care. Of the remaining 1.3 per cent of the population, 11 per cent were imputed as Carers in comparison to 10 per cent among the remainder of the population.

### *Long-term Illness*

There was a 3.9 per cent non-response rate to this question, which asked about any long-term illness, health problem or disability which limited the person's daily activities or the work they could do. 22 per cent of these were imputed as having such a condition in comparison with 18 per cent among the remainder of the population.

#### *Address One Year Ago*

This question had a non-response rate of 4.5 per cent. No usual address was imputed more often than among the remainder of the population, mainly because there was a high rate of non-response for children under 1.

#### *Qualifications*

This topic was covered by two questions, on educational and professional qualifications, which had non-response rates of 6.2 per cent and 17.2 per cent respectively. Where missing, professional qualifications was set to None by an edit rule if the educational qualifications was answered. Professional qualifications was set to missing if educational qualifications was not answered. Taking the responses to the two questions together, a new variable called highest qualification was derived. After applying the edit rules, 6.1 per cent of people needed to have highest qualification imputed. People with imputed values were more likely to have no qualifications (Level 0) than the remainder of the population.

#### *Activity Last Week*

This variable shows whether a person was working in the week prior to Census day, and if not whether they were looking for work, waiting to start a job, retired, student, looking after home/family, permanently sick or disabled, or otherwise economically inactive. This information is derived from Questions 18 to 22 on the Census form for people aged 16 to 74.

Problems were found with the pattern of responses to these and other employment questions which was caused by the format of Question 18 (Last week, were you doing any work). Some people ticked No or multi-ticked this question, but then went on to give details of their present job in answer to Questions 32 to 36. The filter rules were amended to accommodate this pattern so that they were treated as working.

Non-response to working last week was 2.1 per cent. The value was changed in certain cases depending on the pattern of responses to looking for work etc (questions 19-22), ever worked and year last worked (question 23), details of current or last job at questions 25-30 and current job at questions 32-35.

In total, 3.7 per cent of Activity Last Week values were imputed. These were biased towards looking for work and most of the economically inactive categories, especially retired and students. Only 34 per cent were

imputed as working compared with 64 per cent in the remainder of the population aged 16-74. Generally it was people at the extremes of the age range who failed to respond to these questions, which explains the preponderance of retired people and students among the imputed values.

### *Employment Status*

Question 25 asked whether each person who had ever worked was an employee, or self-employed with or without employees in their current or last job. Non-responses and multi-ticks amounted to 6.5 per cent of those who should have answered the question. These all went through imputation, and 'Employee' was imputed more frequently than among the remainder of the population.

### *Size of Workplace*

The non-response rate for this question was 13.9 per cent. An edit rule was applied to set the number of workers to 1-9 where a person was self-employed without employees. This left 6.5 per cent to be imputed, of whom slightly fewer were set to 1-9 workers than among the remainder of the population, and slightly more in the 10-24 and 25-499 ranges.

### *Occupation and Industry*

The non-response rate for occupation was 3.2 per cent among currently working people, including 0.7 per cent inadequately described responses. When all people who had ever worked are considered, non-response rose to 13.1 per cent. The imputed population was slightly biased towards people in major groups 4 (administrative and secretarial), 7 (sales and customer services), 8 (process, plant and machine operatives) and 9 (elementary occupations). Occupation groups 2 (professional) and 3 (associate professional and technical occupations) were under-represented.

A similar pattern can be found in non-response to the question on industry. Non-response was 7.8 per cent among current workers, including 0.6 per cent inadequately described, but reached 17.9 per cent taking into account all people who have worked. Imputation created more people working in sections A (agriculture), F (construction) and O (social and personal services) and fewer in D (manufacturing), J (banking, finance, insurance), L (public administration) and M (education).

It should be noted that the full codes were imputed for missing occupation and industry data. However, the primary matching variables for these fields were defined at the major group level. Thus if industry was reported but occupation was missing, a donor would have been sought within the same major industry group, and that person's

occupation copied into the recipient's record. In some cases an unlikely occupation/industry combination may have been created at the individual code level.

### *Supervisor*

Question 29 asked whether people supervised any other employees in their current or last job. The non-response rate was 6.8 per cent. An edit rule set missing answers to No unless occupation was also missing. This accounted for about half the non-response. Of the remainder, 25 per cent were imputed as supervisors compared with 30 per cent among the remainder of the population.

### *Workplace Address*

There was a 7.8 per cent rate of non-response to this question, but some values could be deduced from the answers to method of travel to work. This left 6.4 per cent to be imputed. Of these, fewer were imputed as working at or from home than amongst the remainder of the population.

### *Method of Travel to Work*

This question was asked only of currently working people. Non-response was 6.3 per cent, which was reduced to 5.0 per cent by a set of edits. The imputed values were biased towards public transport users and those travelling by foot and away from working at/from home or driving a car or van.

### *Hours worked*

The non-response rate was 8.0 per cent, and imputation favoured the 0-19 hours per week range compared to the pattern among the remainder of the population.

## **Household Variables**

	Total (including imputed)	Non- response	Imputed	Non- response	Imputed
	000s	000s	000s	%	%
Accom. type	22,305	671	671	3.01	3.01
Self-contained	22,305	870	870	3.90	3.90
Number of rooms	20,542	1,117	1,116	5.44	5.21
Bath/shower and toilet	20,542	503	503	2.45	2.35
Lowest floor level	22,305	897	919	4.02	4.12

Central heating	20,542	539	442	2.62	2.17
Number of cars	20,542	669	554	3.26	2.72
Tenure	20,542	797	685	3.88	3.36
Landlord	6,582	-	175	-	2.94

Note: The 'Total' column refers to the number of households in scope for the question, i.e.:

- Accommodation type, Self-contained, Lowest floor level: All household spaces whether occupied or not (including enumerators' responses on dummy forms where no Census return was made)
- Number of rooms, Bath/shower and toilet, Central heating, Number of cars, Tenure: All occupied household spaces, i.e. households with at least one usual resident
- Landlord: All occupied household spaces where Tenure was renting or rent free

#### *Accommodation Type*

There was a 3.0 per cent non-response rate for this question, which was asked of all households. Imputed values were more likely to be a purpose-built flat, part of a converted or shared house, or a commercial building, and less likely to be a detached or semi-detached house.

#### *Self-contained*

This question had a non-response rate of 3.9 per cent. Of imputed households, 1.5 per cent were given not self-contained status compared with 1.1 per cent among the remainder of the household population.

#### *Number of Rooms*

Question H3 provided two boxes for the number of rooms occupied by a household, so that any value from 1 to 99 could be entered. Early analysis of processed data showed that there were some problems which needed to be addressed:

- a zero entered into the left hand box was sometimes interpreted by OCR as a 6, creating values from 61 to 69 instead of 1 to 9
- a diagonal slash entered into the left hand box was sometimes mistaken for a 1, creating values from 11 to 19
- if the form-filler attempted to make a change by crossing out and writing a different figure in the other box, both figures might be recognised by OCR or the number could be duplicated in clerical processing. Thus where a value of 3 was changed to 4, the number of rooms might have been interpreted as 33, 34, 43 or 44.

After carrying out an analysis of households with more than 10 rooms, rules were put in place to set values to missing where they were greater than a number which depended on accommodation type. Number of rooms was subsequently imputed. No limit was applied to detached houses.

Imputation was slightly more likely to set a value of 3 or 4 rooms, and less likely to impute 5 or more rooms, compared with the remainder of the household population.

#### *Bath/shower and toilet*

Question H4 asked whether a bath/shower and toilet was available for use only by the household. There was a non-response rate of 2.5 per cent and slightly more households were imputed as lacking sole use than among the remainder of the household population.

#### *Lowest floor level*

4.0 per cent of households failed to answer this question. Fewer were imputed as having ground floor as their lowest level of accommodation than the remainder of the household population.

#### *Central heating*

This question had a non-response rate of 2.6 per cent. Non-respondents were slightly more likely to lack central heating than for the remainder of the household population.

#### *Number of cars*

There was a 3.3 per cent non-response to this question. 35 per cent of these households were imputed as having no cars compared with 26 per cent for the remainder of the household population.

#### *Tenure and Landlord*

Non-response to these questions was 3.9 per cent for tenure and 2.9 per cent for landlord. Those not answering were more likely to be renting and less likely to be outright owners than in the remainder of the population. Among tenants, there was little bias towards any type of landlord among the imputed group.

### **Comparison with 1991**

In general, the biases found in the imputed values for the person and household variables were in the same direction as those present in the 1991 Census data, but were less marked. For example, 52 per cent of those imputed in 1991 for marital status were assigned as Single compared with 41 per cent in the Census population. In 2001 the

corresponding proportions were 49 per cent and 44 per cent. 53 per cent of non-respondents were imputed as having no car in 1991, considerably higher than the 32 per cent among reporting households. In 2001, when the non-response rate had risen from 1.0 per cent to 3.3 per cent, the gap had narrowed to 34 per cent among imputed households and 27 per cent in those who responded.

### **How well did EDIS work?**

Within EDIS, a number of assumptions were based on age being correct rather than other items. However, year of birth was occasionally mis-stated, not scanned correctly or given a wrong value during processing. Particularly when there was an error in the next to last digit of the year, EDIS may have imputed for a range of items where no value was needed, or conversely set reported data to 'no code required'.

Late changes to the questionnaire design had some impact on EDIS. Splitting the qualifications questions into two parts, which occurred after the 1999 Rehearsal, meant that new rules had to be devised for 2001 to deal with professional qualifications. This turned out to be the question having the largest non-response rate as many people considered that it did not apply to them.

As extra room had to be found on the form for qualifications, the question on work last week was squashed into a smaller space. As a result, two of the bullet points which appeared separately on the Rehearsal form were conflated into one, and it appears from the pattern of responses to this and later questions that some form-fillers misunderstood the question and answered No when they were actually in work. A resolution was found to this problem by amending the filter rules for the derivation of Activity Last Week but a small number of answers may have been miscoded as a result of the extra complication which was introduced.

A single edit and imputation system was designed to deal with the censuses in England, Wales, Scotland and Northern Ireland, which all had slightly different requirements. Variations in the design of the Census form and in editing requirements meant that great attention had to be devoted to ensuring that the processing for each country was carried out to the desired standards.

### **Conclusion**

EDIS was successful in its main aim of providing a complete and consistent database of values for all people who completed Census returns. It did so efficiently and largely followed standard principles of making minimum changes to the data. There were complications in its development including late amendments, some of which could have been

avoided with earlier access to live data and others which were due to changes between Rehearsal and the final version of the Census. However, these issues were identified at an early stage of Census processing.

Further results on the performance of EDIS will be reported in the 2001 Census Quality Report, which is due to be published later this year.