



Census 2001 Review and Evaluation

November 2005

2001 Census Samples of Anonymised Records (SARs) Evaluation Report

Content	Page
Introduction	2
Consultation	2
Extraction of Data	3
Small Cell Adjustment.....	3
Release of Data	3
Use of Data	4
Conclusion / Possible Lessons for 2011	5

ONS is carrying out a review and evaluation of the 2001 Census in England and Wales which will culminate in a Data Quality report being published, followed by a General Report.

Plans for individual reports on specific aspects of the Census operation and a timetable for release have been published.

Each report is written in isolation and is subject to amendments as processing progresses and further information comes to light.

Reports will be released on the ONS website in the form of a high level Executive Summary and a more detailed Evaluation Report.

Census Customer Services
ONS
Titchfield
Fareham
Hants PO15 5RR

Telephone: ++44 (0) 1329 813800
Fax: ++44 (0) 1329 813587
Minicom: ++44 (0) 1329 813669
E-mail: census.customerservices@ons.gov.uk
Website: www.statistics.gov.uk/census2001

Census 2001 Review and Evaluation

Aims

The aim of the project was to produce Samples of Anonymised Records (SARs) from the 2001 Census. Licensed samples have been anonymised by removing names, addresses and other information which could lead to the identification of an individual person or household. ONS was commissioned by the Economic and Social Research Council (ESRC) through the Cathie Marsh Centre for Census and Survey Research (CCSR) to deliver the following three products:

Licensed Individual SAR

The Individual SAR consists of a 3 per cent sample of individuals, relating to some 1.76 million people in the UK. For each person it contains the main demographic, health and socio-economic variables and derived variables eg social class; household information; data on sex, economic position and social class of the individual's family head. Geography area identification is at Government Office Region level in England and at country level for Wales, Scotland and Northern Ireland. For details of the specification see www.ccsr.ac.uk/sars/2001/

Special Licence Household SAR

The Household SAR consists of a 1 per cent hierarchical sample of households for England and Wales only. It has around 245,000 household records together with information on individuals within those households. For each household it contains the main variables and derived variables as for the Individual SAR with additional household information such as size of household. For details of the specification see www.ccsr.ac.uk/sars/2001/hhold/index.html

Licensed Small Area Microdata (SAM)

The SAM contains information for the whole of the UK and is a similar dataset to the Individual SAR. It consists of a 5 per cent sample and around 2.9 million individual records with geography area identification at local authority level for England, Wales and Scotland and Parliamentary Constituencies for Northern Ireland. For details of the specification see www.ccsr.ac.uk/sars/2001/

Project Scope

The aims and scope of the project widened as it became apparent that the approach to Census output production and range of disclosure control issues impacted greatly on the timetable of the delivery of the SARs. The concerns that there were a greater array of other sources and greater computing power available, meant that the 2001 licensed SARs would provide less detail than in 1991. The change of environment meant that stronger confidentiality protection was required. ONS recognised this and set up a mechanism whereby applications could be made by researchers to carry out valuable research projects using data with even greater detail than was available in 1991. ONS produced the Controlled Access Microdata Samples (CAMS) which are the in-house versions of the individual and household SAR. The CAMS required significant resources in setting up mechanisms for access and in carrying out a pilot scheme for assessing feasibility of wider research use. The CAMS have minimal disclosure control treatment applied to the data and are only available within the ONS safe setting environment.

Individual Controlled Access Microdata Sample (CAMS)

The Individual CAMS provided in-house is the same 3 per cent sample of individuals as the Individual SAR with full detail for most variables and at much greater detail than the 1991 SARs. Index of multiple deprivation (IMD) scores are available for England and Wales and Northern Ireland. IMD deciles are provided for Scotland. A number of other derived variables are also available such as household membership and family composition. It contains geographical identification to local authority level, apart from the smallest areas (City of London, Isles of Scilly and some Scottish areas). For details of the variables and codes see www.ccsr.ac.uk/sars/2001/indiv-cams/codebook/index.html

Household Controlled Access Microdata Sample (CAMS)

The Household CAMS provided in-house is the same 1 per cent hierarchical sample of households as the Household SAR but also includes records for Scotland and Northern Ireland. It has limited groupings of variables allowing comparability with the publicly available 1991 Household SAR. The same IMD information, other derived variables and geography

Census 2001 Review and Evaluation

detail have been applied to this file as for the Individual CAMS. It also contains additional household derived variables for specific household analysis. For details of the variables and codes see www.ccsr.ac.uk/sars/2001/hhold-cams/codebook/index.html

Background

The case for providing Samples of Anonymised records from the 1991 Census was made by Marsh et al (1991). It was concluded that there were great benefits from having access to Census microdata in a way that also protects confidentiality of respondents. The ESRC through the CCSR commissioned the ONS to produce the 1991 SARs and at the time the risks of disclosure seemed relatively low. Two files were released: 2 per cent samples of individuals with geography at Local authority level set at a population threshold of 120,000 and a 1 per cent hierarchical household file with geography set at Government Office Region. These datasets were widely used by the research community and a large number of innovative and policy-relevant papers were produced. CCSR provide support and training on the SARs and they also modify the file formats so that they are suitable for common statistical software packages.

The ESRC submitted a request for comparable files from the 2001 Census and, in addition, a 5 per cent Small Area Microdata file (SAM), but more geographical detail with a higher level of grouping on individual level information than the Individual SAR. Dale and Elliot (2001) reported a consultation on the user requirements for the 2001 SARs and assessed the disclosure risk using the same methods as had been used for the 1991 SARs. It was concluded by Dale and Elliot that increasing the sample size made only a small increase in risk to confidentiality.

However, ONS had been going through a period of review and evaluation on the importance of protecting confidentiality given that the social and technological environment had changed markedly since a decade ago. It was important to avoid identification and disclosure through linkage with other datasets given the increasing computer power to match datasets and the perceived proliferation of easily accessible databases holding personal data. The National Statistician has a strict obligation not to reveal information collected on individuals in the Census. It is against this background which the protection of the SARs was assessed.

ONS concluded that given the increased risks, statistical methods would be used to measure these risks. Judgements on other factors such as how the data would be used and whether or not some users may have datasets that could be matched against the SARs to identify individuals would also be considered. On this basis, ONS decided that no geography below region would be available on the Individual SAR; that single years of age would not be possible and that a number of other variables would need to be more heavily grouped than in the 1991 SARs.

CCSR and ONS held user consultations with the research community in 2002 on the content of the revised 2001 SARs, following the initial feedback from ONS. These consultations were largely to assess which bandings of variables were most acceptable to users. ONS wanted to provide as much detail as possible for researchers without compromising confidentiality.

Further details on the consultation can be found at www.statistics.gov.uk/census2001/sar_consultation.asp. In order to meet the Census Offices' legal requirements, the level of detail was reduced on a number of highly visible or identifying variables (eg age, occupation, industry, geography) in the SARs. A decision was taken to focus on the Individual SAR and to delay an assessment of the disclosure risk of the Household SAR until the disclosure work on the Individual SAR had been completed.

The scope of the disclosure issues was not recognised during initial planning. This led to inevitable delays to proposed timetables and this was compounded by other Census output production being treated as higher priority. The original timetable provided by Census Division for the three SARs products was for planned delivery to CCSR in summer 2002. This was seen to be extremely optimistic and was revised to April 2003, then September 2003. By December 2003 it seemed clear that it was not appropriate to give a definite timetable for the SARs until two months before each product was due to be released.

The user community expressed concerns over restrictions in variable detail and the continued delays in the SARs production schedule. As a consequence, in April 2004 ONS made arrangements for access to the more detailed SARs datasets in-house for valuable research work. These are now known as the CAMS (see section on Evaluation for further details). However, user access to the CAMS was delayed until August 2004.

Census 2001 Review and Evaluation

Methods and Approach

A key consideration throughout the SARs project has been the protection of confidentiality. This contributed to the significant delays to the production of the SARs as a lot of the work used pioneering and innovative methods to assess levels of disclosure risk. It was important to provide the right methods for protecting high risk records that allowed as much detail as possible for researchers. The Census Offices could not agree on a specification until they were fully confident that the level of detail would protect the individual records within the files. Quality assurance on the disclosure assessments was carried out by the Confidentiality and Privacy Group (CAPRI), a research group which forms part of CCSR, at Manchester University.

Licensed Individual SAR

Preliminary analysis focused on the UK 1991 SARs, as 2001 Census data were not available at the time. This provided the base from which to draw up the initial specification for the 2001 Individual SAR. The disclosure assessment was based on five “intruder scenarios” of risk employed by Elliot and Dale (1998, 1999).

ONS commissioned Mark Elliot of Manchester University to carry out special uniques analysis - ‘a set of computational methods for identifying risky records’ (Elliot and Manning, 2001). The special uniques methodology assigns a score, known as the DIS_SUDA score, to each sample unique identified on the key. Each score tells us, when a unique match was made between a SAR record and a record in an external database, the probability that it is correct. The analysis also provides information about which variables and variable values are contributing most to the overall risk.

The disclosure risk measure used was the percentage of population uniques in the sample for each of the intruder scenarios. The availability of the full census file allowed us to exactly determine records that were both unique in the sample and unique in the population.

Work proceeded rapidly after Dec 2003 when a 2001 sample extract was made available. The first stage of disclosure control on the Individual SAR was an initial special uniques analysis conducted on the extracted Individual SAR and validated by ONS Methodology Division that led to some recoding of variables. A second iteration of the special uniques analysis

prompted further recoding. The remaining high risk records (records which were known to be population unique and had a DIS_SUDA score greater than a specific threshold) were subject to perturbation i.e. modification of the data. The perturbation was added using a modified version of the Post Randomisation Method (PRAM). For further details on the pramming process used, see Bycroft and Merrett (2005). The PRAM methodology developed for the SARs was quality assured by Statistics Netherlands.

The special uniques analysis was undertaken for England and Wales and Northern Ireland data. However, because Scotland had not adjusted small cells in their tabular output there were greater concerns over the safety of SARs in Scotland. To deal with this, GROS produced all the possible 3-ways tables for Scotland and Sam Smith at CCSR searched these for unique values, based on the proposed specification for the Individual SAR. Unique combinations were marked and sent to ONS for perturbation.

Revised Licensed Individual SAR (Version 2)

The disclosure control techniques applied to the SARs were under continual review. In August 2004, when the Individual SAR was near completion and final disclosure controls were being applied, it was decided that of the five scenarios of risk that were assessed, by far the most relevant was the private database cross match scenario. As a result, the Individual SAR was deemed to be over protected and the loss of detail in some variables was unnecessary. At the same time ONS withdrew the religion variable for England and Wales; and GROS withdrew information on religion for Scotland, in light of an uncertain social and political environment and sensitivity among religious groups.

Given that the first version of the Individual SAR was almost ready, ONS decided to release the ‘over protected’ Individual SAR to CCSR in September 2004 with a second version planned for early 2005 containing more detailed information. CCSR consulted users on the proposals for providing more detail in the Individual SAR version 2, eg increasing ethnicity category breakdown from 5 to 16 for England and Wales and the country of birth categories from 7 to 16. A further special uniques analysis was carried out in-house by ONS Methodology Division and a similar process was undertaken on Scotland data. The results were acceptable.

Census 2001 Review and Evaluation

At the SARs Conference in September 2004 users expressed their dissatisfaction with the withdrawal of the religion variable for England and Wales and Scotland (the Northern Ireland religion variable remained in the SAR) and this was reiterated on several occasions during the production of version 2 of the Individual SAR. ONS listened to users concerns and undertook a review in which the National Statistician consulted with leaders of a number of faith groups. As a result, he formally approved its release and the religion variable was then added back into the Individual SAR version 2 for England, Wales and Scotland. Further testing and PRAM adjustments were carried out before its final release to CCSR in March 2005.

Before users could access version 2 of the Individual SAR they had to sign an agreement which ensured that they deleted all copies of version 1. This was necessary to ensure that the perturbation applied to the files was not compromised.

For further details on variables and codes in version 2 see www.ccsr.ac.uk/sars/2001/indiv/index.html

Special Licence Household SAR

At the outset, the plan was to produce a licensed household SAR, accessible under conditions similar to those employed for the Individual SAR. However, because individual information was required together with household information, the combination and inter-relationship between these variables increased the risk of both households (and the individuals within those households) being unique in both the sample and the population. At the time however, it was still considered possible to protect the data sufficiently by recoding and PRAM adjustments.

The first stage of analysis on the household SAR considered possible recodes for age, sex and marital status for different household sizes to reduce the disclosure risk to a level acceptable for release as a non-disclosive file. However it was soon recognised that implementing the necessary level of coding for these variables meant that the file would be of little value to users, largely because age is often the main focus of analyses at different life stages, eg working age, school age or retirement age. Coding by single year of age for all household sizes to meet the requirements of the user community would mean producing a more disclosive household file which could not be released under the current licence arrangements as for the Individual SAR.

Work was done in Spring/Summer of 2005 to investigate the best approach to satisfying customer needs whilst ensuring confidentiality protection and a final specification was agreed based upon the results of the disclosure assessment. In parallel, ONS was 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 methods with ii) legally binding agreements to protect data confidentiality. This approval was extended to the household SAR so that ONS could provide a more useful file accessible via the UK Data Archive through an ONS Special Licence. Despite these arrangements, the smaller size of the sample in Scotland and Northern Ireland meant that the risk of disclosure was too high for these countries. Accordingly, the Household SAR was restricted to England and Wales.

Following this approval, the household SAR became known as the Special Licence Household SAR because of the detail provided and the special terms under which it can be accessed. In addition to the agreed coding of variables, a small amount of perturbation was applied to protect confidentiality. The file was released to CCSR in August 2005.

Licensed Small Area Microdata (SAM)

The SAM file was extracted in August 2004 and a preliminary disclosure risk assessment was conducted by ONS Methodology Division to assess the extent and degree of recoding required given the more detailed geography. Producing a SAM that was useful to researchers was a success. The early stages of discussion on the SAM development were inevitably cautious as the extent of recoding and consequently the effect on data utility was not known. At a late stage in the SAM development, users requested additional variables that were on the Individual SAR and these were incorporated after their disclosure assessment.

Initially, CCSR provided a draft specification and disclosure analysis based on 1991 variables. At that time there was no commitment from ONS to produce a SAM and it was decided that, if the SAM was accepted in principle, a revised specification would be needed based on the parameters agreed following the initial disclosure assessment.

Census 2001 Review and Evaluation

The same risk assessment as for the Licensed Individual SAR was employed, using knowledge of population uniques and special uniques analysis. Similarly to the Licensed Individual SAR, recoding and PRAM perturbation was used to protect the file.

Controlled Access Microdata Samples (CAMS)

The CAMS files were copies of the original extractions from the Census database of the individual and household SAR. A pilot study, commencing in June 2004, was used to assess the feasibility of setting up Census microdata access and facilitate researchers' use of the CAMS datasets under strict security conditions. A similar precedent had already been established in the Business Data Linking Project within ONS. The secure safe setting environment known as the Virtual Microdata Laboratory (VML) was provided in London which facilitated bonafide research on business datasets. ONS established a link from Titchfield to the VML and trialled four researcher visits. The positive outcome initiated the continuation of the CAMS service.

The individual and household CAMS have since been developed and include derived variables such as household composition, family membership, IMD scores including: income, employment and education; and urban and rural indicators. These make the files more useful to researchers as they link policy related variables to Census microdata. A great deal of work has been undertaken by ONS and CCSR to ensure the CAMS files generated the greatest amount of utility and are user friendly and manageable in a range of statistical software packages. This has added considerable value to the CAMS. The individual CAMS file became officially available for use in ONS following the pilot study in December 2004 and the household CAMS became available in July 2005.

Access to the CAMS is through an application to the Census Research Access Board (CRAB), who have delegated authority from the ONS Microdata Release Panel (MRP) to authorise research using the CAMS on behalf of the Registrar General. Similar arrangements exist for Scotland and Northern Ireland. The CRAB, led by ONS, includes representatives from GROS, NISRA and CCSR. The CRAB meets around every 6 weeks to assess applications based on the criteria for approval. If applications are approved the researchers can access the CAMS to carry out statistical analysis in the VML once the necessary conditions of access are met.

Confidentiality is maintained by having a safe setting within ONS, a strict application process, and checking of all outputs prior to release. It is not possible for a researcher to remove any results themselves. They are placed in an area within the VML and can only be removed by designated by ONS staff. Confidentiality guidelines have been made available to researchers to guide them in submitting non-disclosive outputs to speed up efficiency of their output release. Currently the CAMS can be accessed via the VML at all ONS sites: London, Newport, Titchfield and Southport.

For further information on the application process including the terms of reference of CRAB, criteria for approval and confidentiality guidelines, see www.statistics.gov.uk/census2001/sar_cams.asp

Results

A summary of the project outputs and details of their availability are provided below:

Licensed Individual SAR (UK) – September 2004

This file is no longer available from CCSR see www.ccsr.ac.uk/sars/2001/ for access and registration for version 2 of the Licensed Individual SAR.

Revised Licensed Individual SAR (UK), referred to as version 2 – March 2005

This file is available from CCSR via an End User Licence see www.ccsr.ac.uk/sars/2001/ for access and registration.

Licensed Small Area Microdata (UK) – October 2005

This file is available from CCSR via an End User Licence see www.ccsr.ac.uk/sars/2001/ for access and registration.

Individual CAMS (UK) – December 2004

This file is available at ONS in the VML and access is subject to a strict approval process, see www.ccsr.ac.uk/sars/2001/indiv-cams/access/index.html for details on how to apply.

Census 2001 Review and Evaluation

Special Licence Household SAR (England and Wales) – October 2005

This file is available from the UK Data Archive, to holders of an End User Licence and those who have been authorised by ONS for a Special Licence see www.ccsr.ac.uk/sars/2001/hhold/ for details of how to apply for access.

Household CAMS (UK) – July 2005

This file is available at ONS in the VML and access is subject to a strict approval process, see www.ccsr.ac.uk/sars/2001/hhold-cams/ for details on how to apply.

Evaluation

Assessment and lessons learnt

Delivery of Products

The delivery of the SARs files was a major challenge to ONS as the user community had hoped for similar detail to that of the 1991 SARs. In addition, as the SARs products were treated as secondary Census outputs this affected the timetable significantly, particularly in the early stages. The delivery of the SARs files took up a far higher than anticipated level of resource, particularly from ONS Methodology, which far exceeded the planned funding. This was due in part to a major rethink in the disclosure control methodology to be applied. In addition, high staff turnover in all stakeholder areas in ONS has meant that consistency and expertise has not been easily maintained.

There was a delay of several months in recruiting a new project manager which impacted on the overall project timetable. The combination of low priority status and unexpected high resource demand gave rise to major delays in drawing up, agreeing and extracting and protecting the individual and household specifications, which resulted in dissatisfaction within the user community.

Once the SARs data were extracted and received by the contractor working for ONS Methodology Division in December 2003, only then could Methodology carry out the detailed risk assessment. The first special uniques analysis was carried out on the 2001 data and a dataset (Licensed Individual SAR) was constructed and released by September 2004..

The SARs project began to be formally managed by the SARs Project Board which first met in December 2003. Led by ONS, the Project Board consisted of representatives from GROS, NISRA, ONS and CCSR. From that point onwards the project began to take shape and gain momentum. The Project Board was recognised as contributing greatly to the delivery of SARs outputs as it provided the focus for project progress and a forum for decision making and issue resolution. CCSR and the Census Offices recognised the value of the Board's significant contribution to the delivery of the SARs outputs.

The 2001 SARs being given lower priority than many other Census products was one of the main reasons for late delivery. The approach is being reconsidered for Census 2011. The initiation of a new SARs Project Board will be recommended for delivering the SARs in 2011 and it is also recommended that this Board is in place alongside other Census Project Boards well before the 2011 Census, which was not the case in 2001.

Impact of Statistical Disclosure Control assessment

The major challenge in producing the SARs was balancing user needs with requirements for confidentiality protection which have been shaped by the Census Acts and the National Statistics Protocol on Data Access and Confidentiality www.statistics.gov.uk/about/national_statistics/cop/downloads/prot_data_access_confidentiality.pdf

The Census Offices have a clear and well published obligation to protect the confidentiality of information. The general strategy for ensuring the statistical confidentiality of 2001 Census output was stated in the Government's March 1999 White Paper (www.statistics.gov.uk/census2001/pdfs/whitepap.pdf) The 2001 Census of Population (Cm4253):

"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:

Census 2001 Review and Evaluation

- 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 basic head counts – will be amalgamated with that for a significantly large enough neighbouring area; and or
- modifying the data before the statistics are released.”

A huge amount of work was carried out following the consultation in September 2002. Users requested enhancements to the 1991 SARs specification, see paper 'Proposal for 2001 samples of anonymised records: an assessment of disclosure risk' by Dale and Elliot (2001). This paper concluded that the disclosure risk from the 1991 SARs was overestimated and this was partly due to the low probability of being able to cross match records from the 1991 General Household Survey (GHS). For the 2001 SARs there were requests for an increase in the sample size and a decrease in the population threshold.

ONS assessed the disclosure risks not just in terms of quantitative statistical methods but also qualitative assessments given the changed social and political environment and increase in computing power since 1991. The main analysis was to determine whether a variable should be collapsed, the number and proportion of unique individuals and an assessment of the risk that an individual within the SARs could be identified by matching the SARs against an external dataset. Overall the disclosure assessments and agreement on treatments were essential work but commenced too late and took far longer than envisaged at the outset.

As part of the ONS strategy for 2011 Census to include SARs in its output planning, the disclosure control approaches for the 2011 Census will also include consideration of the SARs microdata and should be coordinated with the methods used for tabular output.

Scotland and Northern Ireland

As the intention of the SARs Project was to provide users with UK wide datasets ONS involved the other UK Census Offices on consultations over the SARs products. However, because of their different needs, there are some variables within the specifications that are specific to one or more countries only. This is because there were some different questions and question formats on Census day. It is accepted that there can be different levels of detail in some variables in Northern Ireland and Scotland data compared with England and Wales. See specifications for further detail.

The disclosure control measure known as small cell adjustment was used in the 2001 Census for England and Wales and Northern Ireland tabular output. GROS did not implement small cell adjustment on Scotland data, so there was a higher chance of matching individual SAR records with other census output. In light of this, the disclosure control assessment had to take two different approaches. The special uniques analysis was carried out by Mark Elliot under contract to ONS Methodology Division on England, Wales and Northern Ireland data. This approach was not possible with the data for Scotland and so Sam Smith from CCSR undertook a separate exercise for Scotland which consisted of the identification of population uniques, based on the coding used in the SARs, for 38,000 specially commissioned tables. Unique combinations were marked and returned to ONS for perturbation.

Overall, the disclosure control analysis and treatment required a two pronged co-ordinated approach and contributed to some of the early delays in reaching decisions on how best to assess the risky records on Scotland data. In general the communication between ONS, the other Census Offices and CCSR was not at the level required in the early stages of the project. This resulted in a lack of information about progress which contributed to delays in reaching agreement on issues relating to individual countries. The establishment of the SARs Project Board improved communication significantly and the input from representatives from GROS and NISRA was of great value.

Census 2001 Review and Evaluation

A wider issue, that impacts on the Census output, including the SARs, is the lack of consistency between the three Census Offices in, questions, formats, categories, edit imputation and the different approaches taken to address disclosure risk. It would be a far reaching benefit to UK users, as well as the SARs, if the Census Offices could agree to greater harmonisation in these areas.

Data Quality and Sampling Information

Though the delivered files have proved of considerable benefit to the research community, there have been some problems relating to the variables on the SARs products. Quality assurance does take considerable resource, and can delay delivery, but the long-term benefits may outweigh the short-term delay. There have been some occasions where incorrect variables, format or labelling have caused difficulties for users and further work was required post-release to correct.

There has been a general lack of definitive information documenting the sampling strategy. The effect has impacted on the investigation of the over sampling of the Individual SAR (the file has a sample size of around 3.1 per cent instead of 3 per cent). The lack of definitive documentation showing how this was produced has resulted in difficulties in explaining the over-sampling. The 1 per cent Special Licence Household SAR and the 5 per cent Small Area Microdata have the correct sample sizes.

Licensed Individual SAR

There were further delays to the project outputs as the initial disclosure assessment for the 2001 Individual SAR was based on scenarios of risk that were not formally agreed by the senior policy committee within ONS until a very late stage in the production of the Individual SAR. Alterations in the disclosure control approach meant that the first version of the Individual SAR was overprotected and was a poor quality product unacceptable to users. Rather than delaying delivery of the first version, ONS endeavoured to produce a second version to replace the first, at a later date, once analysis was complete. However, this meant that production of the household SAR stalled as resources were limited. In hindsight, ONS should have approved the scenarios of risk at a much earlier stage; this might have avoided the need for a revised version of the Individual SAR.

Special Licence Household SAR

The ESRC's specification for the Household SAR was too disclosive and resulted in a higher number of population uniques than was acceptable for a typical licensed file. After the disclosure assessment, it appeared not to be possible to produce a household SAR that was both useful to the research community (i.e. based on the categories of the Individual SAR) and suitable for use via an End User Licence. Producing a file that met user requirements achieved confidentiality protection through an appropriate balance between data modifications and licence arrangements was a challenging task for ONS. Extensive analysis on coding options was needed until the right balance was achieved for England and Wales, though this could not be achieved for Scotland or Northern Ireland. Access to the household SAR would be provided using a combination of standard disclosure control methods and a Special Licence, which placed additional restrictions and conditions on the researcher and their institution, with strong sanctions for breach of licence conditions. This was a pioneering and innovative solution for Census microdata.

After it was agreed to produce a more detailed file, the work on the Special Licence Household SAR was completed by ONS in 6 months. The file was then deposited at the UK Data Archive a month later following additional work by CCSR. This was delivered more quickly than the Individual SAR. Though still in its infancy, ONS consider the Special Licence Household SAR a success and has set a precedent for future Census microdata to be accessible using this approach.

Given that the Special Licence Household SAR is available from the UK Data Archive under Special Licence arrangements. ONS are considering how this type of access may be beneficial for Census 2011 microdata.

Licensed Small Area Microdata (SAM)

The case for SAM is detailed in Tranmer et al (2004) who noted that there was huge potential for 'a unique source of geographical and individual level information'. ONS recognised the importance of research at the local level for policy development particularly for use in central and local government. The benefits were seen in terms of supporting Census 2001 standard tables with relevant local authority SARs

Census 2001 Review and Evaluation

data. The request to investigate the production of a SAM file was new and one had not been created for 1991. However, the SAM was similar to the Individual SAR except that it was larger, at 5 per cent, with local authority level detail.

The methodological practices had already been developed for the individual version which enabled a smoother production process for the SAM. Inevitably there was a trade off between the constraints of confidentiality, the amount of individual level detail and geographical breakdown in any file that could be produced.

The initial population uniques analysis that was carried out on the original coding (similar to the Individual SAR) highlighted the need for coarser categories, given a greater sample and more detailed geography. There were doubts that such a file with significantly less detail than the Individual SAR would be useful, which meant that the future of the SAM was uncertain. Users requested that the production of the SAM went ahead and any decisions on its usefulness would be assessed once the disclosure assessment had been completed.

The production of a SAM file was deemed a welcome addition to the SARs outputs by the user community. If, subsequent to release, feedback is generally positive, ONS would consider the option of producing a similar dataset for 2011.

Controlled Access Microdata Samples (CAMS)

The Individual and Household SARs took longer than expected to produce, as a higher level of disclosure control was required than for the 1991 SARs. For example, many variables had been subjected to broader categorisation and the geographical detail for the 2001 SARs was planned to be at regional level. The 1991 SARs was recoded by local authority for those over 120,000 population, whilst smaller LAs were grouped together. In light of this, users argued that 2001 data would not be comparable to 1991 and they wished to access more detailed microdata for research purposes under appropriate conditions.

Supplying more detailed microdata in a safe setting (eg through the VML) was given approval as long as confidentiality could be maintained. These datasets have been well used by researchers and as of 11 November 2005, around 30 applications have been approved for access to the CAMS.

This service arrangement has been praised by the research community because they have access to much more detailed data than was previously available. ONS did not initially plan to provide the CAMS product, but, doing so has provided significant benefits which have had considerable policy development value. Although there is some inconvenience for researchers in needing to travel to an ONS site, and there are improvements to be made to the stability of the VML itself and the associated statistical software packages, the CAMS has been viewed by some researchers who need more detail as a very useful addition to the SARs files.

ONS have noted the value of the research and the importance to users in providing access to detailed Census microdata and are considering the options for providing access to 2011 Census microdata.

Consultation with the user community

RSS SARs Conference – 30 September 2004

SARs User Group Meeting – 15 July 2005

There has been continued user representation from Angela Dale of CCSR on the SARs Project Board, which has meant we have had ongoing dialogue with users. Feedback has been encouraged from users accessing the CAMS and this led to improvements in the CAMS service. For example the extension of the VML to other ONS sites and the CAMS User Guide have made practicalities of access easier and helped users to know what to expect when they come to visit ONS.

Census 2001 Review and Evaluation

The Census Offices have improved their links with the SARs user community and gained important feedback on the SARs and the CAMS. It was at the RSS SARs Conference that the issue of adding the religion variable was first raised and the user reaction was strongly in favour of putting it back in. In light of the general feeling about the removal of this variable and the assurances of confidentiality, the religion variable was included in the revised version of the Individual SAR. This demonstrated that ONS acted on users views and reviewed the decision accordingly.

The SARs User Group meeting highlighted how important access to the CAMS had been, particularly to social research relevant to government policy. The research presented included employment differences by ethnicity, differences in 'head of household' by ethnicity and local authority, analysis of local labour markets, and determinants of internal migration. It was interesting to note that researchers were using CAMS because it provides more detail particularly on age, as single year of age has been provided and smaller geographies. It also complemented and supplemented data not available in standard tables.

There was praise from researchers for the service provided by ONS staff in accessing the CAMS. The greatest concern was the high cost to researchers in terms of time and travelling to access the CAMS at an ONS site. The possibility of remote access to microdata was raised by more than one researcher and this will be considered in parallel with other options for 2011 microdata access. More access points such as in Scotland and Northern Ireland were seen as highly beneficial. In accessing the CAMS there have been problems when the server has been working at full capacity because access is shared with business data users. Steps are being taken within ONS to purchase a second server that is designed to improve response speed.

The panel discussion highlighted a number of areas where ONS could improve the production, dissemination and access to the SARs and CAMS for Census 2011. ONS have listened to these helpful insights and the majority of them are reported in the recommendations below.

Conclusion

The production of the 2001 SARs and CAMS has been a journey of development for ONS, GROS, NISRA and CCSR. Initial plans to produce the Licensed Individual and Household SARs to a specification equivalent to that in 1991 had to be revised due to increased concerns over confidentiality protection and a greatly changed technological environment. Alternative strategies had to be developed and tested as the level of detail was not enough to satisfy the user community. Delays to higher priority Census work resulted in a failure to meet the timetable, since a full analysis of disclosure risk relies on access to a final version of Census data. This was further exacerbated by the necessary level of consultation with users and suppliers on getting UK-wide agreement on disclosure issues and coding options. The SARs development was also hampered by limited resources and staff turnover. Furthermore, the lack of proper project management in the early planning of the SARs led to wildly optimistic assumptions about delivery dates.

However, the investigative work to obtain a household SAR and SAM that was of sufficient detail has been applauded. There were even occasions where it seemed unlikely that these products, at the proposed level of coding to minimise disclosure, could ever be delivered. The consultation with users via CCSR has been hugely important and highly valued and this was strengthened in the last 12 months through open meetings for SARs users.

The Census Offices in partnership with the CCSR have responded to users' views and needs and balanced the constraints of protecting confidentiality to achieve a useful set of licensed SARs datasets. Balancing the requirements stimulated the development of new methods of disclosure control which are now being used in other microdata samples. In addition, new approaches for access to disclosive data provided the driver for facilitating access to more detailed microdata through the CAMS.

Despite the inherent difficulties and setbacks within the project, the 2001 SARs have delivered three highly valuable Census microdata files. For the Individual SAR, the data available under Licence are considerably less than users requested and much less than was released in 1991. An equivalent level of detail was available for

Census 2001 Review and Evaluation

the Household SAR, though under Special Licence conditions. However, CAMS files with much more detail than in 1991 files have also been provided in-house at ONS. Therefore there is a net gain in the amount of data available for researchers although the conditions of access have become necessarily more restrictive in order to allow the greater detail many researchers desired. A great deal of experience, knowledge and insight has been gained in producing and disseminating 2001 SARs and this is currently being fed into discussions for 2011 Census outputs production.

Recommendations

There are a number of recommendations that would improve the production of census sample microdata and how it could be accessed. The key important areas for consideration are outlined below:

- Plan the SARs work thoroughly from the outset, taking into account relative priorities. Understand all the dependencies before suggesting delivery dates. This would have the benefit of giving microdata a predefined priority alongside other Census outputs.
- Confirm both internal and external resources from the outset and establish a SARs Project Board at an early stage to facilitate the development of the SARs and CAMS for Census 2011 microdata.
- The disclosure control methods needed for census tabular data and microdata are interdependent. Therefore, development of census disclosure control strategies should consider delivering methodology in such a way that both tabular output and microdata extracted from it are useful and safe.
- Maintain links with the user community, exchanging views on developments for the 2011 Census, and continue to listen to feedback on the SARs and CAMS datasets already released.
- Determination of acceptable levels of risk should be made early on in the development process and remain firm.
- Incorporate into the project plan a thorough programme of data quality assurance, and ensure that adequate resources are available to support this.
- Continue to provide information on statistical disclosure control. Promote awareness of obligations to protect respondent confidentiality

and the need to employ statistical disclosure control methods. Investigate more fully the effects on data quality and analysis, recognising consequent resource implications.

Consider increasing the sample size of census microdata in particular, the feasibility of providing a larger sample and smaller geographies in the safe setting environment.

- Investigate alternatives to improve the accessibility of CAMs and reduce the need for researchers to visit an ONS site to carry out analysis. Options include: extension of VML access to GROS, NISRA, Universities and ONS Regional Offices (essentially increasing the number of safe-settings); providing a remote access facility; and provision of synthetic data.

References

- Bycroft C. and Merrett K. (2005) Experience of using a Post Randomisation Method at the Office for National Statistics, *United Nations Economic Commission for Europe Work Session on Statistical Data Confidentiality*, Geneva, November 9-11 2005.
- Dale A. and Elliot. M (2001) Proposal for 2001 samples of anonymised records: an assessment of disclosure risk, *Journal of the Royal Statistical Society A*, 164, Part 3, pp 427-447.
- Elliot, M. J., and Dale, A. (1998) Disclosure Risk for Microdata. *Report to the European Union ESP/ 204 62/ DG III*.
- Elliot, M, J., and Dale, A. (1999) Scenarios of attack; the data intruder's perspective on statistical disclosure risk, *Netherlands Official Statistics* Volume 14, spring 1999, special issue: Statistical disclosure Control
- Elliot, M. and A. Manning (2001) "The identification of Special Uniques," *Government Statistical Service Methodology Conference*, June 2001, London.
- Marsh C.; Skinner C.; Arber S.; Penhale B.; Openshaw S.; Hobcraft J.; Lievesley D.; Walford N. (1991) The Case for Samples of Anonymised Records from the 1991 Census, *Journal of the Royal Statistical Society, A*, 154, Part 2, pp 305-340.
- Tranmer, M., M. Brown, A. Dale, M. Elliot, E. Fieldhouse, C. Gardiner, D. Martin, A. Pickles and D. Steel (2004) "The Case for Small Area Microdata," *Journal of the Royal Statistical Society series A* 168, 1.

Census 2001 Review and Evaluation

Census Topics	Target Dates for Release
Legislation	Published
Non-Compliance (Executive Summary Only)	Published
Data Needs	Published
Geography	Published (Executive Summary)
Publicity	Published
Data Collection Development	Published
Data Collection Support	Published
Census Coverage Survey	Published
Processing	Published
Annex: Quality of Data Capture and Coding	Published
Downstream Processing	Published (Executive Summary)
Data Quality	
- Question non-response rates	Published
- Disclosure Control (Executive Summary only)	Published
- Data Validation (Executive Summary only)	Published
Edit & Imputation	Published
One Number Census	
- Quality Assurance	Published
- Lessons learnt (Executive Summary only)	Published
Output Policy	Published (Executive Summary)
Output Production	
- Part 1:Review of Output Released to date	Published (Executive Summary)
- Part 2:including Sample of Anonymised Records (SARs)/Origin Destination Matrices	Published
Census Access	Published
Programme Management	Published (Executive Summary)
Quality Report	Published
General Report	Published

Please note that the dates for release of individual evaluation reports noted above are target dates, and therefore subject to change. For the latest information please visit www.statistics.gov.uk/census2001/reviewevaluation.asp