
SARS NEWSLETTER

NO. 20 – October, 2003

Samples of Anonymised Records from the 1991 and 2001 Census

Census Microdata Unit Faculty of Social Sciences and Law University of Manchester Manchester M13 9PL

THE SARS FOR 2001

This newsletter is designed to update SAR users on recent developments in the production of the SARs. Unlike the usual Newsletters which contain contributions from different team members, this Newsletter focuses on a single topic and is written by Angela Dale.

Timetable for the SARs

We have just heard from ONS that the Individual SAR will not be available until spring 2004. ONS are still firmly committed to producing SARs and have undertaken to give two months' notice of the delivery date. However, there is no firm timetable and it is unlikely that the data will be sent to CCSR before April 2004. The most recent specification is available from the SARS website: <http://www.ccsr.ac.uk/sars/2001/request/>. Note that there are changes to details for religion in Northern Ireland.

The Household SAR will follow after the Individual SAR. A revised specification will be available from the CCSR website by the end of October 2003.

All details in the specification are subject to revision following a further round of confidentiality checks during December-February.

There is likely to be a Small Area Microdata file at a 5% sample and geography at LA level. As soon as we have a specification, we will put it on the web.

More details are available in the following pages.

The production of the SARs 2001

Angela Dale, CCSR

The context

In summer 2000, after extensive consultation with users, CCSR provided a specification for the SARs to ONS and in September 2001 we sent a consolidated specification with all relevant documents attached (<http://www.ccsr.ac.uk/sars/2001/request/>). This specification requested a 3% Individual file, a 1% Household file and a 5% Small Area Microdata file. We agreed that the production of the Individual and Household files should take priority over the Small Area Microdata.

The original timetable assumed the SARs specification would be finalised (i.e. all details agreed by ONS) by September 2002, so that production could start immediately thereafter with delivery in summer 2003. (This timetable would be similar to that for 1991 where the data was delivered to CCSR in August 1993.) However, it was only in October 2002 that ONS responded to our original specification. The ONS response proposed a considerable reduction in the detail of the files and resulted in a further consultation exercise. CCSR held meetings in London and Manchester to obtain the views of the user community and written responses were sent to ONS and placed on the CCSR website. The responses made clear the extent to which the reduction in detail would jeopardise research plans.

In February 2003 I wrote to Len Cook, the national statistician, expressing my concern over the delays in producing the SARs. In reply I was told that 'it is still our intention to deliver the SARs in September'. ONS planned to finalise the specification by the end of April 2003, with the extracted data being subject to the special uniques analysis during summer of 2003.

The current situation

The planned timetable has not been met. We are now in a situation where the extract of the 3% Individual SAR is timetabled for mid-November. This file will conform to the specification on the CCSR/SARs website. There will be no geography below region (except an Inner/Outer London distinction) and we do not yet know whether ONS will agree to add area-level descriptors. If agreed, these would be added after the release of the main file.

Once the file has been extracted from the census database, a 'special uniques' analysis will be conducted with results going to ONS in early 2004. This analysis will identify the level of 'risk' in the file and the variables that contribute most to that risk. On the basis of this assessment ONS will decide whether further recoding of variables is needed or whether the risk identified can be dealt with by perturbing specific variables in specific records. There is clearly a balance to strike between these two forms of protection, both in terms of the integrity of the data and the length of time this process will take.

We have asked ONS to answer the following questions:

- What criteria will be used to decide that additional global recoding is needed?
- What methods will be used to perturb risky records?
- How will the resulting bias be assessed?
- What is the timetable for this work?

As the answers to these questions become available they will be published on the ONS website, as long as this is not felt to jeopardise confidentiality. At the moment, we know that it is unlikely that the Individual SAR will be available before April but we have no firm release date. There is also a strong possibility that there will be a reduction of detail by comparison with the specification on the CCSR website.

The specification for the Household SAR should be agreed with ONS by the end of October 2003 and will then be placed on the CCSR website. Once the data has been extracted this file, too, will be subject to further assessment of disclosure risk. ONS will, as for the Individual SAR, decide how to deal with this risk. This could, for example, result in loss of the household matrix which provides information about each individual to each other in the household. We do not have a date for delivery of the Household SAR.

The special case of Scotland

The concern over the confidentiality of microdata relates in large part to the ability to match to other records in the public domain and thus identify an individual or reveal further information about them. One of the areas of concern relates to the ability to match the microdata records to individuals who are unique in the aggregate statistics. In theory, this could allow an intruder to reveal more geographical information than released on the SAR. However, the decision by ONS and NISRA to round small numbers in output tables has removed this risk for England and Wales and Northern Ireland. Scotland decided not to follow this route and to allow unique values in table cells.

For this reason, special precautions are needed before a safe SAR can be released for Scotland. It is necessary to ensure not only that there is no possibility of making a unique match between existing tables and the proposed SAR but that *there is no possibility of such a match with any future commissioned tables.*

To address this we have obtained from GRO(S) all the possible 3-way (and thus also 2-way) tables at the Scotland level - 38,000 in total. We are now assessing the number of unique cells in these tables and the variables that contribute most to them. We will then assess the extent of additional recoding or perturbation that would be needed to ensure a SAR where no individual could be matched with a unique value in the tables for Scotland. We are not yet in a position to provide the answer to this - but it may mean that the SARs for Scotland have less detail than those for England and Wales. However, in the meantime these tables can be made available to any users who would find them of value.

How to proceed from here?

The requirement to protect confidentiality

The Census Offices have a statutory requirement to preserve the confidentiality of responses to the census and any breach of confidentiality would be highly damaging. We live in a world where individual information is becoming more readily available, computing power is increasing rapidly and search engines are becoming ever more sophisticated. In this situation the Census Offices have a primary responsibility to ensure the safety of any data released from the Census. We recognise and respect that.

A framework for considering confidentiality

When assessing how to ensure confidentiality, there is a recognition that the level of protection built into the data must relate to the level of risk. Data in the public domain - for example, data that can be freely accessed from the ONS website - must be safe to all foreseeable attempts to breach confidentiality. Thus, great care has been taken to ensure the safety of the census area statistics. With microdata files there is a recognised distinction between 'safe data' and a 'safe setting'. Microdata that is so safe that it can go into the public domain with little, if any, restriction may have lost much of its research value. This is a prospect we are currently facing with the 2001 SARs. By contrast, data that is held within an entirely safe-setting (e.g. within ONS) and where outputs can be tightly controlled, can be much less heavily protected and can therefore retain a great deal of detail. For example, the ONS Longitudinal Study contains very detailed individual information but is kept under very secure access conditions and all outputs are screened for disclosure risk.

However, between these two polar opposites of 'safe data' and 'safe setting' may lie intermediate positions where the level of safety of the data can be adjusted to the risk level of the setting. For example, microdata used in a research setting lends itself to institutional and individual safeguards that are entirely consistent with good research practice. This allows one to define 'research' files that contain somewhat more detail (and less perturbation) than the public use versions but where there is a concomitant increase in security. For example, before releasing the 1991 SARs, the ESRC required all academic institutions where staff or students wished to access the data to take legal responsibility for ensuring the confidentiality of the data. All academic users were also required to sign confidentiality undertakings and use of the data was regularly monitored by CCSR. Imposing formalised and legally binding institutional and individual agreements provides much greater control of risk than is typical with 'public use' files.

For access to maximum detail in microdata, it will always be necessary to implement a safe-setting. However, this has very substantial costs (both for the institution providing access and the researcher using the facilities) and has been shown to generate a much lower level of research than microdata that is released (Watkins and Boyko, 2002). Whilst safe-settings have a role to play - for example with the ONS Longitudinal Study - they cannot substitute for access to released microdata files for research purposes.

How does this apply to the SARs?

The extensive work on disclosure control currently being undertaken on the SARs is designed to reduce to a minimum any risk that an individual could be identified. It will provide a very high level of protection. This has involved developing new methods to identify risk in the SARs and this, in turn, has required new procedures and methods for dealing with the risky records so identified. This has two consequences. It is time-consuming and, because the methods are not tried and tested, it is difficult to assess the length of time that will be taken. This also means that there has been no assessment of their impact on the integrity of the data. Loss of detail in the data (through recoding) and bias introduced through perturbation may reduce the research value of the data to the point where they are no longer fit for purpose. These factors, therefore, have led to delayed production and concerns that the final product may have only limited research value.

In the following section there are four recommendations that I am making to the Office for National Statistics. I would greatly welcome the views of users on these recommendations.

1. The need for to recognise a 'research' version of the SARs

The SARs are designed to be a research tool. The 1991 SARs have been used for a large amount of important research. A summary of the key research results using the 1991 SARs can be downloaded from the CCSR website and will shortly be published in *Sociology* (Li, 2003). If the 2001 SARs are unable to deliver the same level of research, this will be a huge loss - not just to the academic world but also to policy makers who make use of the research results. Earlier paragraphs have outlined the kinds of institutional safeguards that can be used to ensure that the data are used only for research purposes where attempts to identify individuals play no role.

Recommendation 1: that ONS urgently implement a 'research' version of the SARs that recognises the safeguards that can be implemented through an institutional setting.

2. What can be achieved through technological advancements?

ONS now have in place the technical facility to allow analysis of microdata on a 'safe' files server inside ONS. So far this has been planned for use with business data (for example the New Earnings Survey) where legal restrictions prevent the release of microdata files. This facility needs to be extended to the SARs so that detail that has now been lost from the 2001 SARs - e.g. local authority geography and the most detailed occupational information - can be analysed in this safe setting.

Recommendation 2: that ONS allow access to a more detailed version of the SARs in a safe-setting.

3. How to help users who have immediate and pressing research commitments to use the SARs?

Given the unexpected delay to the 2001 SARs, there are strong and urgent arguments for allowing in-house, safe access to the SAR file that will be extracted in mid-November. For this file, all the agreed derivations will have been completed and all the standard anonymisation will have taken place. However, any additional disclosure that will be required following the 'special uniques' checks will not have been implemented. Therefore the additional recodes or perturbations that are likely to reduce the research value of the data will not have been conducted.

Recommendation 3: that ONS provide in-house access to the Individual SAR as soon as it is extracted (mid-November) for those researchers who have an immediate and pressing need.

4. A longer-term perspective

There is value in exploring how technical developments can improve the security of microdata whilst also ensuring ease of access. For example, there may be scope to build on the 'safe-setting' facility that is now being implemented in ONS to extend it to allow access remotely from other designated sites or from designated computers. For example, a university should be able to provide a safe-setting with the necessary guarantees to allow remote access. The benefits of the Grid and other e-science developments are likely to be able to play a role in providing easier access to high quality research data whilst also ensuring the confidentiality of the data.

Recommendation 4: that ONS and ESRC work together to explore the scope for harnessing e-science and Grid-related developments to secure remote access to microdata.

References

- W. Watkins and E. Boyko (2002) Safe Data, Safe Places: not either/or solutions, paper presented to CEIES conference on 'Innovative solutions to providing access to microdata', Lisbon. September 2002
Y. Li (2003) The Samples of Anonymised records for social science, *Sociology* (forthcoming)

Meeting with ONS

ONS have agreed to hold an urgent meeting to discuss our concerns over the 2001 SARs and I will be taking forward these recommendations. At the moment they apply only to England and Wales and would need to be separately negotiated with GRO(S) and NISRA.

Your views are important. **Please let me know, as soon as possible:**

1. Your views on the recommendations, above, that we are putting to ONS
2. Whether you are currently waiting for the SARs and the urgency of your requirement
3. Information on research you had planned but may not now be able to conduct
4. Any other views

Please send to: angela.dale@man.ac.uk or write or phone using the contact points below. If you do not already subscribe to the SARs mail list, please go to <http://www.jiscmail.ac.uk/lists/sars.html> and join. We can then make sure we keep you up to date with events.

SARs Contact Details

Website: <http://www.ccsr.ac.uk/sars/>
SARs Helpline (0161) 275 4735

Email: sars-helpdesk@man.ac.uk
Fax: (0161) 275 4722

Professor Angela Dale
(0161) 275 4876
Angela.Dale@man.ac.uk