Disclosure Control

Project Objective

1. To have a series of measures in place that will uphold the 2001 Census confidentiality commitments that published tabulations and abstracts of statistical data do not reveal any information about identifiable individuals or households.

Background

2. The General Register Office for Scotland (GROS) has made commitments publicly to ensure the protection of information collected from the 2001 Census. One commitment was given on the Scottish Census form:

   The information you provide is protected by law and treated in strict confidence. The information is only used for statistical purposes. Anyone using or disclosing Census information improperly will be liable to prosecution. The information on your Census form will be treated as confidential for a period of 100 years

   Similar statements were made on the forms used by the Census Offices for other parts of the UK:

- The Office for National Statistics (ONS) responsible for the Census in England and Wales
- The Northern Ireland Statistics and Research Agency (NISRA)

3. Two years before the Census (and before devolution) the UK Government published the White Paper *The 2001 Census of Population* (March 1999) with the following:

   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.*

4. A robust disclosure control strategy is essential for respondents to cooperate with the Census and other government surveys and ensures that GROS has the highest possible quality statistics available. GROS has to implement procedures that will protect information within all census output. This includes output in tabular form and within microdata (e.g. Samples of Anonymised Records). This paper describes the procedures, but

does not evaluate the statistical attributes of disclosure protection (about which the Census Offices commissioned a number of reports from Professor Chris Skinner and colleagues at Southampton University).

Disclosure Protection in 1991

5. In 1991, the disclosure of information in census output was protected by:

- minimum population thresholds of tables;

  (i) 16 households and 50 persons for Small Area Statistics (the equivalent of the Census Area Statistics for 2001)

  (ii) 320 households and 1,000 persons for Local Base Statistics (the equivalent of the Standard Tables for 2001)

- the design of tables where data were presented for small geographical areas, categories for some variables were often banded to protect the data; and

- the technique of Cell Perturbation. This method introduced uncertainty into published census output based on 100% of Census records by modifying cell counts by up to +2 and -2 in published tables. Unfortunately, a consequence of the Cell Perturbation method was that there were some inconsistencies of data within, and between, tables, through a loss of additivity. (A substantial part of output – that based on 10% of records processed for 'hard-to-code' topics – was not subjected to Cell perturbation on the grounds that the sampling of records provided protection.)

Methodology

6. Most of the work developing a disclosure control strategy for the 2001 Census was done by ONS. GROS conducted research to evaluate the impact of various proposals from ONS on Scottish output. The ONS programme consisted of four main elements:

   (i) A review of 1991 disclosure methods and increased risk of disclosure since 1991.

   (ii) A programme that researched possible disclosure options leading to proposals.

   (iii) An external review of proposed procedures.

   (iv) Further consideration of risk and revised proposals.

A review of disclosure control for the 1991 Census and increased risk since 1991

7. The review of the 1991 methodology concluded that it effectively protected information about identifiable persons and households and that at least the same level of protection should be provided in 2001. The table design and population thresholds worked well in 1991. However, an alternative tabulation method was required due to inconsistencies that appeared in tabular output as a consequence of Cell Perturbation. The Census Offices also

needed to consider options of protection that would address any increased risk of disclosure since 1991 as a consequence of improvements in technology and the impact this has on the availability of data and the ease with which an intruder may identify individual information.

8. The review identified a number of risks, of which the following were relevant to Scottish output:

- The 2001 Census results would be very widely disseminated via the internet. This means that users and the general public can acquire census data more readily and easily than ever before. The increased accessibility also increases the risk of misuse of census data.

- Census data users can obtain large volumes of census statistics freely and we would need to manage increased risk from attempts to break any confidentiality protection provided.

- All questions from the 2001 Census would be fully coded. Previous censuses had coded only 10% of the responses for some key variables and that had added a level of uncertainty to published results.

9. ONS explored each disclosure option following a set of criteria:

- The effectiveness of the method for disclosure protection

- The impact of the method on the quality of census data

- The practical aspects of implementing the method

10. The programme investigated an alternative to Cell Perturbation that could be applied to the Census database before tabulation and thereby provide tables without inconsistencies. The following options were investigated:

- record swapping - swapping a household record with a similar record in the same geographic area;

- data switching - swapping the values of one or more variables in one record with the values for the same variables in another record; and

- over-imputation - randomly deleting variables in existing records and imputing the variables using the Edit and Donor Imputation System.

  Record swapping was chosen by ONS as it was easily implemented and did not substantially damage the quality of the data. GROS preferred over-imputation and asked a potential supplier of software for edit and imputation to include this method in their proposals. In the event, the supplier's price was too high, and software for edit and imputation – and record swapping - was written in-house by ONS.

11. By August 2000, in a UK discussion paper on the Census Area Statistics, the UK Census Offices announced they were planning to use the following disclosure control methods:

- Setting a target or average size for output areas; some 50 households in Scotland as in 1991, though the other Census Offices would adopt a target of around 120 households.

- Setting a minimum size of areas for key output; e.g. 20 households and 50 residents for Census Area Statistics (CAS), and 400 households and 1000 residents for Standard Tables.

- In Scotland, creating only one set of output areas; two sets of overlapping output areas (OAs) could be 'differenced' to create unintended below-threshold areas.

- Limiting the detail in classifications used in tables.

- Record swapping before tabulation.

External review

12. This approach was independently reviewed by Dick Carter of Statistics Canada. He completed his report in September 2000. It stated that he 'was of the opinion that the planned disclosure control strategy is, subject to qualifications ..., appropriate to safeguard the confidentiality of the respondents' information in tabulations produced from the 2001 Census'. Recommendations dealing with these qualifications were accepted by the Census Offices in a statement by the Registrars General in March 2001 (ISBN 1 85774 437 3). In the statement the RsG say "We have accepted all the recommendations and have acted on them accordingly or are taking action to ensure that they are implemented."

Further consideration of risk and revised proposals

13. Before the statement by the Registrars General was published, ONS decided to conduct more research to explore possible options for addressing the increased risks since 1991. GROS carried out various analyses to assess the impact on Scottish output of the various options being considered by ONS.

14. First, ONS looked at the minimum size of areas for output. They confirmed that the design of tables and the population thresholds (in 1991) were effective measures that did not affect the quality of the data and could be repeated in 2001. However, ONS considered that the increased risk of disclosure meant that the population thresholds would need to be increased – although Carter had not made any such suggestion. After analysing the impact on Scottish output, GROS decided that increasing thresholds for the Census Area Statistics (and hence the minimum size of OA) did not offer increased protection commensurate with the damage done to the continuity between 1991 and 2001 OAs. For example, the proportion of the non-White population in very small numbers in OAs decreased very little even when thresholds were doubled.

15. Second, ONS concluded that record swapping had limitations because it would not be apparent to a person using the Census data that any methods of disclosure protection had been implemented. There would be a perception that persons and households were identifiable (particularly for a single count) and the observer might act upon the information as if were true. If counts of 1 and other small values appeared in tables, then

there would be a perception that ONS had not done all that it could to fulfil its legal obligations of confidentiality and thus had not ensured that all possible steps were taken to prevent inadvertent disclosure. There was, therefore, a requirement for disclosure measures that made persons and households with unique characteristics within an area invisible in tabular output. Therefore two options of further protection based upon cell count modification were also considered:

- Rounding of all counts to a multiple of 3
- Small Cell Adjustment (SCA).

16. Users were consulted in 2001 about these options for the tabular modification of cell counts. The issue was controversial and a large number of users preferred to have no additional disclosure protection measures. Where users indicated a preference, small cell adjustment was the preferred choice. This was largely due to the advantage that the method allowed tables to be internally additive and only adjusted small cells. The disadvantage of the method was that knowledge of the adjustment method and comparing adjusted version of the same number in different tables carried the risk of allowing that number to be deduced.

17. Except for 5 tables on workplace populations (see later in paragraph), GROS decided not to use SCA on basis of view about users' perceptions. If it is believed that the user sees 1s and 2s in the cells of a table as disclosive, then it may be decided to introduce SCA because it removes them. But it is possible that the user will see 0s as disclosive - in that a row or column with 0s in all but one cell appears to be disclosive – and SCA increases the number of 0s and hence increases the likelihood of leaving a row or column with one non-zero cell. GROS believed that either approach to the perception issue could be supported and were not convinced that there was a case for changing earlier decisions. Moreover, whatever the perception, tables without SCA based on place of residence are not actually disclosive, because of record swapping. But because record swapping does not alter place of work in a record, SCA was applied to some tables on workplace.

18. ONS and NISRA decided to use SCA, having taken the different view from GROS about users' perceptions.

Measures finally implemented in Scotland

19. Disclosure of information in Census output for Scotland is prevented by the following combination of methods. All three UK Census offices use all methods A to F but not always in the same way or to the same extent. The corresponding information for the rest of the UK is in square brackets.

    A. Setting a target or average size for output areas (50 households). [120 households]

    B. Setting a minimum size of areas for key output (e.g. 20 households and 50 residents for Census Area Statistics). [40 households and 100 residents]

C. Creating only one set of output areas (two sets of overlapping output areas could be 'differenced' to create unintended below - threshold areas). [ONS were planning for the possibility of more than one set of OAs, but so far have only issued one. NISRA are currently in the process of finalising the 2001 Census Grid Square product which they are aiming to release in June 05. Detailed work was undertaken to assess the disclosure risk associated with differencing this output with that already available for Output Areas in Northern Ireland and it was judged that the risk was negligible. The work undertaken by NISRA was independently reviewed and endorsed by leading academics in the field.]

D. Limiting the detail in classifications used in tables. [No major differences - many tables in common]

E. Record swapping before tabulation with rates of swapping dependent on 'noise' already in data introduced by imputation, etc. [Standard rates throughout E&W that did not depend on the 'noise' already in the data].

F. Small Cell Adjustment (workplace tables). [All tables]

20. Methods A to D are aimed at ensuring that there is only a limited number of cases in which all the households or persons in one of the categories of a variable in a table belong to a single category in another variable. When this happens, information can be disclosed from the table about those households or persons. For example, if there were only one Chinese person in an output area, a table for that output area tabulating ethnicity (with 'Chinese' as a category) by employment status would reveal that person's employment status. Therefore, a further measure is needed so that no one can be certain that any such instance relates to actual individuals or households. That measure is Method E which completes the disclosure control package by swapping a small proportion of Census records. A swapped record is then tabulated in a different output area from where the data was collected. This means that a singleton in a table can never be taken as relating to a known individual or household but the aggregated statistics are not materially affected.

21. A small number of tables have been subject to Method F whereby cells containing small numbers were adjusted randomly. These tables are those on OAs as place of workplace rather than place of residence. Methods B & E are effective for data on residence rather than workplace.

22. The possibility of disclosure is also reduced by the degree of error in Census information (as there would be in any large-scale data collection exercise). This applies particularly to counts of people travelling to place of work or study by area of destination or counts of migrants by area of previous address. The accuracy of such counts cannot be controlled in the same way as counts by area of residence (at the time of Census). General information about data quality may be found at Census Update no. 19

23. In addition to the above, one of the conditions of using Census data is that users will undertake not to attempt to obtain or derive information about specific individual or household, not to claim to have obtained or derived such information.

Assessment and Lessons Learnt

24. The main lesson learnt from the project were that some elements of the disclosure risk assessment should have been carried out much earlier. It was less than one year before Census day, in 2000, that ONS concluded that they would need to take extra precautions to protect information as a result of the increased risk since 1991.

25. The method of perturbing the Census database before tabulation created a few unforeseen problems:

- GROS, without small cell adjustment, have had to take a different approach from the other UK Census Offices to controlling the disclosure risk for samples of anonymised records.

- The effect of record swapping is to alter the place of residence of selected households without altering any other characteristics of these households and the people in them. A swapped housed – and each person in it – is tabulated in an output area of residence other than the one in which it was enumerated. But there is no change in a person's output area of travel destination or migration origin. Thus record swapping is ineffective when counting records against travel destination or migration origin. While these two items are based on postcodes of travel destination and address one year before the Census and are subject to relatively high levels of error, GROS decided to use small cell adjustment for certain workplace tables.

  Compared with 'over-imputation', record swapping cannot be targeted to particular variables, nor is it very effective with records for large households and for people in communal establishments, where pairs of records for swapping may be hard to find.

26. The requirement that all output for England and Wales and Northern Ireland was to be subject to small cell adjustment meant that

- Each Census Office had to maintain its own database and this introduced complexities into producing output based on data from more than one country. Such output included not only tables for the UK but also tables on migration and travel for single countries.

- In making contributions to Scottish tables on migration and travel, ONS and NISRA had to produce separately the output for each type of area (council area, health board, etc). Without small cell adjustment, only output for Scottish Output Areas would have been needed as GROS could have aggregated it as required to higher areas.

Conclusions

27. The Disclosure Control project has, with some qualifications, achieved its objective. Each disclosure measure alone would not provide adequate protection, but the combination of all measures provides sufficient protection to meet the commitments to protect confidentiality. The record swapping, small cell adjustment and threshold constraints have been successfully implemented.

28. The disclosure control project aimed to design and implement procedures to protect information within all census output and this aim has been achieved. The project began with a review of the procedures for disclosure protection in 1991, identifying problems that occurred with these methods and assessing increased risk resulting from the increased use of electronic resources. The review showed that an alternative to Cell Perturbation was needed, and record swapping was implemented in its place. In 2000, the overall approach was endorsed by external review. The three UK Census Offices subsequently adopted different packages following a late re-assessment by ONS and NISRA of disclosure risk.

29. Valuable lessons have been learnt, mainly about the timing of the assessment of disclosure risk and the timing of consultation with users. With this knowledge, and the successful implementation of the measures, there is a good basis on which to build future disclosure control strategies. Final decisions should be reached much earlier for a future census – and there are advantages in the 3 Census Offices adopting the same policy. Over-imputation should be considered afresh in the light of the shortcomings of record swapping.