Scotland's Census 2011
Shaping our future

Paper 2 SCSC 2010/11

Statistical Disclosure Control Update

June 2010

Table of Contents

Statistical Disclosure Control Update

1 Purpose

1.1 This paper is to inform the Scottish Census Steering Committee (SCSC) about the plans for protecting personal data in the published outputs from the 2011 Census. It gives a brief description of the chosen method and sets out the work required to develop this into a fit-for-purpose system for 2011.

2 Background

2.1 The census is unique in the amount of detail that it reveals about the population, both in terms of the large number of variables available and the small populations that can be studied. At the same time, it is important that information about particular individuals and households does not become public. Besides our legal and moral duty to protect personal information, we also have to consider the effect on the response rate if we could not provide a guarantee of confidentiality.

2.2 Although most census data is published in the form of tables, it may nevertheless be possible to deduce certain pieces of information about some individuals. Figure 1 demonstrates how this may occur:

Figure 1: Sample table showing disclosure

	All people	White	Indian	Pakistani and Other South Asian	Chinese	Other
Total persons	**87**	**79**	**3**	**0**	**1**	**4**
No qualifications	**11**	10	0	0	1	0
Group 1	**13**	10	2	0	0	1
Group 2	**10**	10	0	0	0	0
Group 3	**10**	9	0	0	0	1
Group 4	**39**	37	1	0	0	1
Not aged 16 to 74	**4**	3	0	0	0	1

2.3 In this case, there is only one Chinese person in the area and it is easy to tell that this person has no qualifications. A row or column with only one non-zero cell, such as the ‘Chinese’ column in this case, is always disclosive. Some instances of disclosure are less obvious: for instance, in this case a user who is, or knows, the one Indian with a group 4 qualification, can tell that both the other Indians in the area have only group 1 qualifications. This is because they can mentally remove the cell containing the person that they know, which once again leaves a column with only one non-zero cell.

2.4 To entirely remove the potential for such disclosure would mean severely restricting the amount of data that can be published, which would be unacceptable to users and defeat the purpose of the census. Instead of this, it is normal practice in censuses worldwide to use statistical methods -

collectively known as statistical disclosure control (SDC) - to create doubt as to whether a particular published value is true. Of course, these methods should not damage the data to the extent that it is no longer fit for purpose - a balance needs to be struck between the risk of disclosure and the utility of the data.

2.5 In 2011, following extensive research into a number of options, the three UK census offices have jointly decided to use a method known as targeted record swapping. This is a refined version of the method used by General Register Office of Scotland (GROS) in 2001.

3 Description of basic method

3.1 Record swapping is a pre-tabular method of SDC, which means that all adjustments are made to the data before any tables are produced. This has a number of advantages over post-tabular methods (such as rounding), including:

- all tables are guaranteed to be consistent
- tables are quicker to generate, as SDC does not have to be run each time
- the software used to run SDC does not need to be retained indefinitely

3.2 As the name implies, the method involves swapping a proportion of records between geographical areas. Usually a household will be paired with a very similar household in an adjacent output area. This means that variable distributions at output area level will be unaffected, or in a few cases changed only slightly; and that distributions at higher geographical levels, such as local authority, will be entirely unchanged.

3.3 It should be noted that record swapping does not necessarily remove disclosive cells - in fact, in some cases it will apparently create new ones. Protection is provided by creating doubt in the mind of the user as to whether a particular cell value is genuine.

3.4 Figure 2 shows the disclosive table from figure 1, after some record swapping has been applied. In this case two individuals have been swapped with a similar person from a neighbouring output area: the Chinese person with no qualifications has been replaced by someone of the same ethnicity but with a group 1 qualification, while the Indian person with a group 4 qualification has been replaced by another Indian, who has only a group 1 qualification.

Figure 2: Sample table following record swapping.

	All people	White	Indian	Pakistani and Other South Asian	Chinese	Other
Total persons	**87**	**79**	**3**	**0**	**1**	**4**
No qualifications	**10**	10	0	0	0	0
Group 1	**15**	10	3	0	1	1
Group 2	**10**	10	0	0	0	0
Group 3	**10**	9	0	0	0	1

Group 4	**38**	37	0	0	0	1
Not aged 16 to 74	**4**	3	0	0	0	1

3.5 It can be seen that there is still only one Chinese person, so a user who is unaware that SDC has been applied may wrongly assume that the one Chinese person in the output area has a group one qualification. The column headed 'Indian' also now has only one non-zero cell, so it now appears - again, wrongly - that all the Indian people in this area have a group 1 qualification. As long as the user is aware that some SDC has been applied, doubt is now created as to whether the facts apparently revealed about these individuals are actually true - even if these cells had not in fact been affected.

3.6 In this case, individuals were swapped with someone of the same ethnicity. This means that the marginal totals of the ethnicity variable are unchanged, although the distribution of the qualifications variable has been altered slightly. The total number of individuals in the table will never be changed by record swapping.

4 Refinements

4.1 In 2001, each record had an equal probability of being swapped. While this was found in an independent review[1] to provide sufficient protection against disclosure, it is possible to refine the method to provide enhanced protection for those records at most risk of disclosure, while minimising the amount of perturbation to data that is unlikely to be disclosive. This is achieved by targeting the most unusual records so that they are more likely to be swapped.

4.2 For example, a household containing two adults and two children, all of them of white Scottish ethnicity, is probably similar to many surrounding households, and so the likelihood of any published table revealing information about this household is low. All records must have some chance of being swapped, so that there is always doubt created as to the true value, but in this case the record need only have a very low probability of being swapped with one from an adjacent output area.

4.3 On the other hand, a household consisting of a single mother of black ethnicity, with two children, one of whom is black and the other of mixed ethnicity, is likely (depending on the area) to be very rare, and it may well be possible to obtain sensitive information about the household from published tables. This record would need to have a higher probability of being swapped. Note that the actual swap rates have yet to be determined, but will not in any case be made public as such knowledge could assist someone who wishes to circumvent the disclosure control system and discover personal information.

[1] *Statistical Confidentiality Review for the 2001 Census of Population*, R. G. Carter (Statistics Canada), 2000

4.4 Some types of record, such as a sixteen-year-old widow, may be so rare that swapping between adjacent output areas will not provide any protection - their details will still be disclosed when tables are produced at a higher geography, such as local authority level. In these cases, a limited amount of swapping across local authority or other higher-level boundaries may take place.

4.5 Other factors, besides the rarity of a record, may influence its probability of being swapped. For instance, an output area with a high proportion of imputed records is likely to be given a lower swap rate, as a degree of 'noise' has already been introduced to the data.

5 Further work

5.1 The continuing development work on the targeted record swapping system includes research into:

- optimum swap rates to use
- method of determining risk level for a particular record
- which variables to use to identify 'pairs' for swapping

5.2 Most of the research is being carried out by the Office for National Statistics (ONS) on behalf of all three UK census offices. GROS has significant input into the work via the UK Census Statistical Disclosure Control Working Group, and is involved in all major decisions. In addition some work by our consultant, Frank Thomas, has fed into ONS's research. In particular, Frank has carried out a large part of the development work on the algorithm to determine which records are most at risk of disclosure.

5.3 Other ongoing work includes investigating disclosure control methods for:

- origin-destination tables (which are not protected by record swapping)
- communal establishment data
- microdata

5.4 Work is also taking place to determine whether the introduction of hypercubes (which allow for more detail and flexibility of outputs than in previous censuses) will lead to a requirement for additional protection. Any such protection would probably take the form of restrictions on the amount of detail that can be obtained at low geographical levels, rather than any additional perturbations (pre-tabular or post-tabular) to the data.

6 Summary

6.1 The primary method of disclosure control for the 2011 Census will be targeted record swapping. This is similar to the method used by GROS in 2001, but with some added refinements. In 2011 the same method will be used by all three UK census offices. Research is ongoing to ensure that personal information is protected without unduly damaging the data or restricting its usefulness.