

Statistical Disclosure Control for the 2011 UK Census

Jane Longhurst, Nicola Tromans and Caroline Young – Office for National Statistics

1. Introduction

UK Census data is released in a number of different formats; standard pre-planned tables, commissioned tables requested by users and sample microdata. In addition in 2011 the aim is to release user defined tables via flexible table generating web-based software. Publishing aggregate or individual data carries the risk that individuals or entities could be identified and confidential information about them could be released.

The aim of Statistical Disclosure Control (SDC) is to ensure that statistical outputs provide as much value to the users while protecting the confidentiality of information. SDC methods modify, summarise or perturb the data and there are a range of different methods that can be used to protect different outputs. This paper describes the strategy that is being employed to develop an SDC solution for UK 2011 Census tabular outputs. The key aim is to ensure a harmonised UK SDC strategy which ensures that the public interest in the figures is met while managing data confidentiality risks.

2. Approach

In November 2006 the UK SDC Policy position for the 2011 Census was agreed by the Registrars General of Scotland, England and Wales and Northern Ireland (ONS 2006). A UK SDC working group consisting of representatives from all three UK Census Offices has been set up to take forward this work and develop a 2011 Census UK SDC Strategy that is in line with the agreed policy. A work plan for the methodological research phase of the 2011 Census UK SDC strategy has been developed. The plan addresses all census outputs whilst taking into account the impact of interactions between them. The initial stage of methodological research involved conducting a review of SDC in a census context considering previous research, evaluations undertaken for 2001 and international practices. Following this a high level review has been conducted to address the advantages and disadvantages of a wide range of SDC techniques. Using this high level review, a preliminary list of SDC techniques which should be explored further will be drawn up. The short-listed methods will be evaluated using a disclosure risk - data utility framework (Shlomo and Young 2006). An additional stage of research will be timetabled to further develop methods for safeguarding microdata. It is vital that the development of the 2011 Census UK SDC strategy takes account of interdependencies such as downstream processing, geography and outputs design.

3. SDC methods

This paper focuses on SDC methods for protecting census tabular outputs rather than microdata samples although the dependencies between the methods used to protect different outputs will be recognised in the evaluation stage. SDC methods for census tables include both pre-tabular and post-tabular methods or combinations of both. Pre-tabular methods are implemented on the microdata prior to the tabulation and typically include forms of record swapping (Willenborg and de Waal, 2001) that can be generalised into PRAM (the Post-Randomisation Method) (Gouweleeux et al 1998). Over-imputation can also be considered, this involves imputing values for randomly deleted variables. Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of rounding. Other post-tabular methods include cell suppression or some form of random perturbation on the internal cells of the tables. The ABS have developed a new cell perturbation method for their 2006 Census. Disclosure risk can also be managed by restricting the design/complexity of the tables, setting geographical thresholds or implementing rules that determine the sparsity of tables or by using access agreements or licences. This paper focuses on a comparison of record swapping and the ABS cell perturbation method in order to demonstrate the evaluation that will be undertaken for SDC methods for 2011 Census.

3.1. Record Swapping

Record swapping involves exchanging geographical variables between randomly selected pairs of households within the census data. In order to minimise bias pairs of households are determined which match on some control variables. Record swapping can be targeted to high-risk households and can also be modified to take into account imputation rates. For this analysis, both random (non-targeted) record swapping and targeted record swapping were carried out for a 10% swapping rate. The control variables that were used to determine the pairs of households were the number, age and sex of persons in the household (using three broad age groups) and a "hard-to-count" index of the household based on the 1991 UK Census enumeration. The record swapping was carried out within a large geographical area (Local Authority (LA)) and households were swapped in and out of small geographical areas (Output Areas (OA)). In addition, targeted record swapping was also implemented by focusing on households that have at least one person in a small cell in one of the tables under evaluation. The main advantages and disadvantages of record swapping are outlined below.

Advantages	Disadvantages
<ul style="list-style-type: none"> ○ Consistent totals between tables ○ Tables are additive ○ Protection offered to both tabular outputs and microdata, (further protection may be required for microdata) ○ Marginal distributions preserved at a higher geographical level and within control strata ○ Some protection against disclosure by differencing ○ Control variables and swapping rates are flexible ○ Swapping geographies will not necessarily result in inconsistent and illogical records ○ Flexible table generation is possible 	<ul style="list-style-type: none"> ○ Effects of perturbation hidden and are hard to measure and account for in statistical analysis ○ Table not visibly perturbed – clear explanations needed to ensure transparency ○ Geographic fields such as workplace are not swapped hence origin-destination tables not protected ○ Method introduces bias ○ In general swapping has attenuating effects whereby associations between variables are lowered and distributions are flattened

3.2 ABS Cell Perturbation

For the protection of their 2006 Census outputs, the ABS conducted research into a new cell perturbation algorithm (Fraser and Wooton 2006). In the past they have released static tables of data however flexible table generation will be used for 2006. The new perturbation algorithm is designed to protect these tables by potentially altering every cell in every table by a small amount. The algorithm always randomises the same table in exactly the same way.

The SDC algorithm involves two stages. The first adds perturbations to the cell counts dependent on the original cell value and a random value or 'cell key'. After the perturbation stage, the same cell in different tables is consistent (has the same perturbation added) due to the use of cell keys which are calculated from the set of records comprising each cell. However, the tables do not add up. Additivity is restored in the second stage using an iterative algorithm which visits single and pairs of cells adding -1,0,+1 at each iteration stopping when all rows and columns add up. It does this at the same time as minimising the overall difference between the additive and original table. A look-up table (that determines the perturbations using the original cell value and cell key) can be designed according to the specification of the statistical agency. For this analysis only the perturbation stage of the ABS method has been implemented since the code for the additivity stage is not currently available. The following look-up table was used:

Original Cell value	Perturbation drawn from the following distribution (based on the random cell key)
0	Remain as zeros
1	Normal distribution with mean 0 and variance 2 truncated at -1 and +5
2	Normal distribution with mean 0 and variance 2 truncated at -2 and +5
3	Normal distribution with mean 0 and variance 2 truncated at -3 and +5
4	Normal distribution with mean 0 and variance 2 truncated at -4 and +5
5+	Normal distribution with mean 0 and variance 2 truncated at -5 and +5

The main advantages and disadvantages of this method are outlined below.

Advantages	Disadvantages
<ul style="list-style-type: none"> ○ Tables are consistent ○ Provides protection for flexible tables ○ Efficient - allegedly has a quick run time ○ Protects against differencing ○ Depending on the design of the look-up table, the method can perturb distributions that are approximately unbiased with small variances ○ Able to produce perturbations for large high dimensional hierarchical or cross classified tables ○ Method is extremely flexible; look-up table can be specifically designed to suit needs and different look-up tables could potentially be used for different tables. Moreover the look-up table could be designed to mimic random rounding for example. 	<ul style="list-style-type: none"> ○ Tables not additive (additivity module is not applied here) ○ Once additivity is applied, consistency is lost ○ Needs to be applied to each table separately ○ Public perception that no disclosure control has been applied (unless incorporated into look-up table) ○ No protection for microdata ○ Method less transparent than others e.g. rounding ○ Depends on the appropriate choice of look-up table which may not be suitable for all tables (i.e. sparse) ○ Statistical effects are highly dependent on the choice of look-up table

4. Short-listing SDC methods for quantitative evaluation using a Disclosure Risk - Data Utility Framework

The quantitative risk-utility framework being used to evaluate the SDC methods is not sufficient on its own. Many SDC methods have qualities which cannot be accounted for quantitatively and thus the advantages and disadvantages of SDC methods must also be addressed. These in combination with the results from the risk-utility assessment will inform the recommended approach to SDC in 2011. The qualitative characteristics that will be considered include the overall practicality and feasibility of implementation, the interaction between different types of output, user acceptance of chosen methods, whether the methods are suitable for flexible table generation, offer

any protection to microdata or protection in terms of perceived disclosure. The advantages and disadvantages of the two SDC methods focused on in this paper were addressed in Section 3. Many SDC methods are feasible for the 2011 Census but all have their own limitations. A high level review of SDC methods will be used to determine a shortlist of methods for further evaluation.

5. Quantitative Analysis of Proposed SDC methods

Once agreed the short-list of SDC methods will be evaluated quantitatively focusing on an assessment of the statistical impact of the method on data utility and disclosure risk. A software package (Shlomo and Young, 2006) developed to calculate a variety of information loss metrics (by comparing the protected data with the original pre-disclosure controlled data) will be used for this analysis. Here we present a limited selection of the information loss measures and one risk measure described in Shlomo and Young, 2006 and use them to compare the two methods for one example table. It should be noted that these are preliminary results and are included as an illustration of the analysis that will be undertaken. A more thorough analysis using a wide range of tables, other risk and information loss measures and investigating more methods and different parameters (e.g. swapping rates, look-up table) will be required for the final analysis.

5.1. Risk and Utility Measures

Disclosure Risk

Let R_i represent the record i , I the indicator function having a value of 1 if true and 0 if false, C_1 the set of cells with a value of 1, C_2 the set of cells with a value of 2, $|C_1 \cup C_2|$ the number of cells with a value of 1 or 2. The disclosure risk measure can be interpreted as the percentage of records in small cells that have not been perturbed:

$$DR = \frac{\sum_{i \in C_1 \cup C_2} I(R_i \text{ not perturbed or imputed})}{|C_1 \cup C_2|}$$

Distance Metrics on Internal Cells of the Tables

Distance metrics are used to measure distortion to distributions. A distance metric is calculated for each row in the table and then the overall average across all of the rows is taken as the information loss measure. This format is used since the rows of census tables commonly represent a geographical area while the columns typically define the categories of a specific table, such as sex \times age group \times economic activity. Standard errors are calculated for each distance metric.

Let D^k represent a row k of table D and let $D^k(c)$ be the cell frequency c in the table. Let n_r be the number of rows in the table. *Pert* refers to the disclosure-protected table and *orig* to the original table. The distance metric considered here is:

Hellinger's Distance:

$$HD(D_{pert}, D_{orig}) = \frac{1}{n_r} \sum_{k=1}^{n_r} \sqrt{\sum_{c \in k} \frac{1}{2} (\sqrt{D_{pert}^k(c)} - \sqrt{D_{orig}^k(c)})^2}$$

Distance metrics can also be calculated for sub-totals and totals of the tables.

Change to Rank Orderings

Changes to the underlying ordering of cell counts within the table can be studied. The original counts are sorted within variable categories (i.e. columns) according to their size and deciles (10 equal groupings) $v^{orig}(c)$ are defined. This is repeated for the perturbed cell counts which are sorted according to both their size and the original order in order to maintain consistency for the tied variables. Deciles $v^{pert}(c)$ are then defined for the perturbed variable after the sort. The information loss measure is the percent of cells that have changed deciles:

$$RC = \frac{100 \times \sum_{c \in k} I(v_k^{orig} \neq v_k^{pert})}{n_k} \quad \text{where } I \text{ is the indicator function and is 1 if the statement is true and 0 otherwise, and } n_k \text{ is the number of cells.}$$

5.2. Data

The effects of the SDC methods will be considered for a table at the Output Area (OA) level for an Estimation Area in England relating to Southampton, Eastleigh and Test Valley. Geography is represented as rows in the table and the other variables span the columns. Table 1 describes the structure of the table.

Table 1: Example table

	Variables and Number of Categories	Number of Persons in the Table	Number of Internal Cells	Average Cell Size	Number of Zeros	Number of Small Cells
Table A	Religion (9) Age-Sex (6) OA (1,487)	437,744	80,298	5.45	47,433 (59.1%)	10,137 (12.6%)

5.3. Results

Table 2: Disclosure risk measures

	10% Swap	Random	10% Swap	Targeted	ABS Perturbation	Cell
Probability that a record in a small cell has not been perturbed	0.651		0.506		0.188	

Table 2 displays the risk measure for the different SDC methods for the two tables. The risk is far smaller for the ABS method in comparison to record swapping because, in this case, there is a higher probability that a small cell would receive a non-zero perturbation. Of course, this result is highly dependent on the look-up table (described in section 3.2). The targeted swap focuses on perturbing small cells and hence the risk is less than for the random swapping method. Other risk measures will be considered in future work.

Table 3: Example Information loss measures

	10% Swap	Random	10% Swap	Targeted	ABS Perturbation	Cell
Hellingers' Distance	1.2875 (0.0249)		1.6027 (0.0265)		1.7388 (0.0228)	
Cells moved into different percentile (groups of 10)	26%		34%		20%	

The ABS method has the highest Hellingers' Distance because of the high probability of small cells being perturbed (see result in table 2). Hellingers' is strongly influenced by change to small cells. The best method (in terms of distortions to distributions) for this table and this metric is random swapping. The targeted swap distorts the distributions in the table more than the random swap as expected. The second test shows how swapping and cell perturbation distort the underlying patterns in the data by changing the rank order of cells. There is a lot of distortion with all three methods because more than 70% of cells have values less than 3 (see table 1).

6. Conclusions

This paper has described the approach that will be adopted to develop an SDC strategy for all 2011 Census outputs. A review of past work (particularly undertaken for 2001) has been conducted and has informed further stages of the project. A high level review of SDC methods has been conducted and will be used to develop a shortlist of methods for further evaluation. Examples of this high level review and a quantitative evaluation (measuring risk and information loss) have been presented for three different SDC methods; random record swapping, targeted record swapping and a cell perturbation method. These preliminary results are included as an illustration of the final more detailed evaluation that will be undertaken. It is recognised that developing a 2011 UK SDC strategy which satisfies all user requirements whilst maintaining a high level of data utility is likely to be an unachievable task hence compromises will need to be made. The final recommended approach to SDC for 2011 Census will be informed by both quantitative and qualitative evaluation and the trade-offs between the different methods will need to be communicated to users.

References

Fraser, B. and Wooton, J. (2006) A proposed method for confidentialising tabular output to protect against differencing, Internal report, Data Access and Confidentiality Methodology Unit, ABS.

Gouweleeuw, J., P. Kooiman, L.C.R.J. Willenborg and P.P. De Wolf (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, 14, pp. 463-478.

ONS (2006) <http://www.statistics.gov.uk/census/2011Census/producingdata/outputconfidentiality.asp>

Shlomo, N. and Young, C. (2006) Statistical Disclosure Control Methods through a Risk - Utility Framework: Proceedings of the Privacy in Statistical Databases CENEX-SDC Project International Conference, Rome, 13-15 Dec 2006.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control, Lecture Notes in Statistics*, 155 , Springer Verlag, New York.

