Evaluating the Short-listed SDC Methods for Census 2011: Interim Report

2nd May 2008

Caroline Young (ONS), Keith Spicer (ONS), Jane Longhurst (ONS) with contributions from Philip Lowthian (ONS) and Natalie Shlomo (consultant from Southampton University).

Management Summary

Work has begun to evaluate the three short-listed SDC methods for disclosure control of tabular outputs from Census 2011. This report provides preliminary results for Record Swapping and Over-imputation at 2, 10 and 20% perturbation levels, presenting their disclosure risk- data utility impact on four different, 2001 Census tables. As yet, no results are available for the ABS cell perturbation method. The objective of this evaluation is to observe the broad statistical effects of the SDC methods to reveal any adverse impacts that may be considered unacceptable to census users. Disclosure risk results for imputation and swapping were similar for comparable levels of perturbation. Imputation of age gave additional protection to tables containing this variable but at the expense of lower data utility. In general, both methods have a 'homogenising' effect as geography is swapped or imputed within LADs. Imputed variables are in effect deleted from the data and replicated from the remaining records which leads to significantly decreased association between variables and a reduction in variance. In contrast swapping distorts the links between geography and the attribute variables, but totals, subtotals and household locations are preserved. Origin Destination tables are severely distorted by all three SDC methods, particularly over-imputation and lead to inconsistencies in the characteristics of the flows. In summary, the results show that, for a similar level of protection against disclosure, over-imputation causes data to lose significantly greater utility than does swapping, whether random or targeted. Hence we recommend that over-imputation be dropped from the short-list. Moreover, we recommend that a different approach to O/D tables, particularly for smaller geographies, be considered such as releasing under licence or access agreement, due to their sparsity.

1. Background

Work is underway to develop a strategy for disclosure control of the 2011 Census. A short-list of SDC methods has been finalised (please see the report by Miller et al. 2007) and these are now undergoing quantitative evaluation. Formal quality assurance has been provided for this shortlist by the UK Census Design and Methodology Advisory Committee (UKCDMAC), and individual Census Project Boards in the UK countries have been consulted, prior to formal sign-off by the UK Census Committee (UKCC).

The short-list was created by assessing a number of SDC methods against a set of criteria that were in line with the policy statement made by the Registrars General. The criteria were split into primary and secondary criteria and an additional requirement was that any method that did not meet one of the primary criteria was not considered for short-listing. Following this assessment four short-listed SDC methods were chosen:

- Record Swapping
- Over-Imputation
- ABS Cell Perturbation Method (developed by the Australian Bureau of Statistics)
- Small Cell Adjustment (SCA) with Record Swapping (included to provide a comparison with 2001)

This interim report describes some preliminary results from the on-going quantitative evaluation. Many features of the analysis can be varied including:

- the level of perturbation applied,
- whether perturbed records are selected at random or from the population of high risk records (i.e. a random or targeted approach),
- the area of study,
- the census tables analysed,
- the risk and utility measures chosen for assessment.

The quantitative evaluation can potentially be very time and resource intensive. **Therefore the objective of this evaluation will be to broadly assess the statistical effects of the methods (e.g. does one method inflate variance while the other has little impact) as well as the general implications for disclosure risk, in order to discount any of the short-listed approaches.** Due to time and resources we are unable to take into account all the features as listed. This interim report will show some preliminary results for over-imputation and record swapping only (comparing to the 2001 benchmark approach). Progress is still being made evaluating ABS cell perturbation and this will be discussed in a later, final report.

2. Data for Analysis

To carry out the disclosure risk – data utility analysis, we obtained unperturbed 2001 Census microdata from different Estimation Areas (EAs) of the UK. Here we show results for one EA: SJ (Southampton, Eastleigh, Test Valley) consisting of 437,744 persons and 182,337 households. Communal establishments are not included in this analysis. For this EA, we have currently analysed four census tables (as proposed in the Miller et al. (2007) report). The number of categories per variable are in parentheses:

(Table 1) ROWS: Country of Birth[1] (2) by Sex (2) by Religion (8)
COLUMNS: Geography (described later)

(Table 2) ROWS: Density of persons in household[2] (4) by Accommodation Type[3] (3)
COLUMNS: Geography (described later)

(Table 3) ROWS: Age (16) by Sex (2) by Marital Status (2)
COLUMNS: Geography (described later)

(Table 4) Origin-Destination (O/D) Table: Cells indicate flows between small area geographies.
ROWS: The origin: where the respondent lives (from SJ EA only)
COLUMNS: The destination: where the respondent travels to work (all England and Wales OAs)
(A cell count of one would imply one respondent travelling from the origin in the corresponding row to work in the destination in the corresponding column. More detail on this type of table can be found in Appendix: section A3.2)

The microdata were perturbed according to the record swapping scenarios (random and targeted) and imputation scenarios (random and targeted) and then tabulated. Small cell adjustment was further applied in the case of random record swapping to simulate the 2001 procedure. The methodology for these approaches is described fully in the Appendix (sections A2.1- A2.3). The methods are broadly comparable in terms of level of perturbation as will be explained. At this stage, records that may already have some protection because they were made up of missing values, and so were imputed, are not considered in this analysis as being any different from the other microdata records (i.e. we assume there has been no imputation for non-response). This is to keep the analysis simple for this evaluation. However the existing protection from imputation may be taken into account in the final recommendation (please note this will not be possible with a post-tabular method such as ABS cell perturbation).

Over-imputation was carried out on census data by the Edit and Imputation branch at ONS using CANCEIS (a specially designed package developed by Statistics Canada to impute missing values arising from item non-response). The set of data used for assessing over-imputation and record swapping were slightly different. The former, referred to as CPCD data, is partially edited census data which was prepared for use in the development of CANCEIS. The latter, referred to as ORCD data, is raw census data (the ORCD data will later be used for carrying out ABS cell perturbation). Table A illustrates how these two datasets differ.

Table A: Differences between CPCD and ORCD datasets

	CPCD (used to carry out over-imputation)	**ORCD (used to carry out record swapping)**
Household Types	Only households containing 1-9 persons but this omits very few households (see corresponding box for ORCD →).	All household types and all household sizes of 1-16 (less than 0.05% of households have more than 9 persons).
Geography (both relate to England & Wales only)	Address, Enumeration Districts (EDs) and above (no Output Areas - OAs). Geographies have a slightly different definition (e.g. CPCD wards are defined slightly differently to ORCD wards).	Address, Postcodes, EDs, OAs, Local Authority Districts (LADs), wards.
Variables on file	Limited number of variables available (however *address* can be used to match geography from ORCD file, before imputation).	All variables available

[1] Country of Birth has two (broad-banded) categories which are UK or non-UK.
[2] Number of persons divided by number of rooms, broad-banded
[3] House, flat or other (e.g. caravan)

For this reason, when assessing the census tables in terms of disclosure risk and data utility, the original (unperturbed) tables used for comparison will differ slightly. Please note that record swapping had already been carried out using the ORCD file, and OAs rather than EDs were used in this analysis. Results after assessing risk and utility on the following four census tables using the following geographies will be shown in this report:

(Table 1) Country of Birth (2) by Sex (2) by Religion (8) by ward (70 – using ORCD, 55 – using CPCD)

(Table 2) Number of persons in household (4) by Accommodation Type (3) by OA / ED (1487 OAs – using ORCD, 903 EDs – using CPCD)

(Table 3) Age (16) by Sex (2) by Marital Status (2) by OA / ED (1487 OAs – using ORCD, 903 EDs – using CPCD)

(Table 4) Flows from OA (1487) to TTWOA (7222) - using ORCD, and from ED (903) to TTWOA (7222) - using CPCD where TTWOA is travel to work OA for all in England and Wales.

Appendix A1.1 provides some summary statistics describing the tables. Despite the differences between the CPCD and ORCD files, our objective is to assess the broad statistical effects of the methods (i.e. does one method reduce level of association between variables and the other not impact on level of association at all) as well as the general implications for disclosure risk, rather than comparing like for like. However the comparability of the tables is something to bear in mind when interpreting results. Preliminary results are discussed in section 4. Not all are included in this report; only the most relevant. Conclusions and recommendations are contained in section 5.

3. Short-listed Methods

- **Over-Imputation**

Over-imputation is relatively unknown and there is no recognised methodology. A new approach was developed which can be found in the Appendix (section A2.1). The variables age and geography were imputed. Random and targeted approaches were performed equivalent to 2%, 10% and 20% perturbation levels.

- **Record Swapping**

Record Swapping was used in the 2001 Census and has been previously tested on 2001 census data by Natalie Shlomo. It involved swapping similarly-paired households (based on control strata) in different OAs within the same LAD. A full methodological description can be found in the Appendix (section A2.2). Random and targeted approaches were performed equivalent to 2%, 10% and 20% perturbation levels.

- **ABS Cell Perturbation**

ABS Cell Perturbation is the final SDC method to be considered for Census 2011. This method is post-tabular whereby table cell values have a perturbation added, drawn from a look-up table. The perturbations in the look-up table are dependent on the original cell value as well as the particular combination of records used to compose the cell. Progress is still being made on implementation of this method so results will not be included in this interim report. Three look-up tables are planned (that control the perturbation added to the census tables); these will result in approximate perturbation levels equivalent to a 2%, 10%, 20% pre-tabular approach.

- **Small Cell Adjustment (SCA)**

SCA is applied to the tables derived from the random record swapped microdata. SCA involves randomly rounding each small cell but for confidentiality reasons, full details cannot be divulged here.

4. Summary of Disclosure Risk-Data Utility Analysis

This section will provide some preliminary results based on the four census tables described. A selection of disclosure risk measures have been developed to analyse whether the short-listed SDC methods give protection against identity, group, within-group and negative attribute disclosure. The Infoloss software (created by Shlomo and Young, 2006) outputs many different results to allow measurement of utility. Results on disclosure risk and data utility are fully contained within the Appendix (sections A3.1 and A3.2). Since the fourth table, which is an Origin-Destination table, is extremely sparse with most of the cells being zeros, risk and utility measures will be used which focuses on the non-zero cells. This fourth table is treated in a separate section in 4.3.

4.1 Summary of Results on Disclosure Risk for Tables 1 to 3

Five measures of disclosure risk have been created to assess the level of protection provided by the SDC methods (as proposed in the Miller et al. (2007) report). The measures are as follows and are explained in the Appendix along with full results.

(i) Identity disclosure
(ii) Group disclosure
(iii) Negative attribute disclosure
(iv) Probability that a small cell value is changed
(v) Within-group disclosure (not yet evaluated)

} explained in section A3.1

The results for the 20% and 10% swaps only will be shown in Appendix section A3.1. We note that table 3 includes age and we might expect this table to result in lower risk for imputation, since both age and geography were imputed. The disclosure risk results can be summarised as follows:

- There appears to be a lot more variability in the disclosure risk outcome of imputation as opposed to swapping, as indicated by table B.

Table B: Comparing the Risk Levels using Measure (iv) after Swapping and Imputation (results show range for targeted and random approaches respectively)

	Table 1	Table 2	Table 3
20% swapping	0.68-0.72	0.74-0.77	0.72-0.72
20% imputation	0.51-0.80	0.70-0.75	0.52-0.58
10% swapping	0.81-0.84	0.84-0.85	0.84-0.85
10% imputation	0.70-0.90	0.87-0.88	0.69-0.74

- A similar pattern is indicated by measure (i); there is greater range in the risk outcome with imputation but overall for tables 1 and 2 these ranges overlap somewhat so we cannot conclude that one approach is better than the other. Table 3 however shows clearly that the disclosure risk is lower for imputation than swapping, as would be expected because table 3 includes *age* which is an imputed variable.

- The effect of SCA with swapping is highlighted in tables 2 and 3 as both methods together result in fewer overall cases of group disclosure (swapping alone shows several cases of group disclosure). There were no cases of group disclosure in table 1. As would be expected, SCA eliminates all cases of identity disclosure as indicated in all three tables.

- Both targeted swapping and imputation tend to reduce the number of cases of *group* disclosure compared to a random approach.

4.2 Summary of Results on Data Utility for Tables 1 to 3

Full numerical results based on the Infoloss software are contained within the Appendix (A3.2) along with a summary of the impact of the methods on the different measures. The formulae for the utility measures can be found in the Miller et al. (2007). An overall evaluation is presented in table C considering the general statistical effects of the methods (which tend to be the same for random and targeted approaches).

Table C: Overall Evaluation of Effects of Record Swapping, Swapping with SCA and Over-Imputation on Utility

Utility Measure	**Record Swapping**	**Record Swapping with SCA**	**Over-Imputation**
Totals of households and persons by geography	PRESERVED Swapping essentially involves moving households around within their LAD. For every household that is swapped out of an OA, one is swapped back in. Thus totals of households and persons are maintained.	SOME DISTORTION This is dependent on the outcome of the SCA (in a large table the number of cells rounded up and down should balance out). *Note that the outcome of SCA can be forced to give exact totals if required.*	NOT PRESERVED Imputation essentially involves blanking out values in a record and replacing with values from a donor record in the microdata. Since geography is one of the variables imputed, extra households could be created or removed. Only EDs are imputed (within LADs), so at the LAD level and above, totals are preserved.
Subtotals of the variables	APPROXIMATELY PRESERVED When aggregating changes to subtotals up to the EA, these changes should balance out as swapping does not actually change the attribute data, only the geographies.	SOME DISTORTION This is dependent on the outcome of the SCA (in a large table the number of cells rounded up and down should balance out).	NOT PRESERVED Since values are being imputed, the integrity of the attribute data as a whole is not preserved, reflected in very large changes to the subtotals (particularly if the table is composed of many categories within variables).
Variable distributions at higher level geography	PRESERVED Since swapping was carried out within LADs, these are maintained.	SOME DISTORTION Due to the aggregate effect of SCA.	NOT PRESERVED Age was not imputed within LAD, and geographies are deleted and replaced with remaining donor values.
Level of association between variables in the table (Cramer's V)	DECREASED SLIGHTLY Since similar households are paired for swapping using control strata, the degree of distortion is small. Amongst the control variables, the level of association is preserved.	INCREASED SLIGHTLY SCA tends to have an opposing effect to swapping.	DECREASED BY A LARGE AMOUNT Over-imputation in general leads the data to become more homogeneous because values are replicated. The effect is particularly significant for higher imputation levels.
Impact on individual cell values (Distance Metrics)	BETTER PRESERVED FOR RANDOM SWAPPING More distortion occurs when unique records are targeted for swapping.	LARGER CHANGES THAN FOR SWAPPING SCA increases the level of distortion in addition to swapping.	BETTER PRESERVED FOR RANDOM IMPUTATION More distortion occurs when unique records are targeted for imputation.
Change to Row Variance	TEND TO DECREASE SLIGHTLY Swapping flattens out the distribution of the cell counts as it is carried out within LADs.	APPROXIMATELY PRESERVED Swapping and SCA have opposing effects on variance.	DECREASED BY A LARGE AMOUNT Imputation flattens out the distribution of the cell counts, but generally more so than swapping as the existing data is being replicated for every imputation.
Rankings by geography	DEPENDS ON LEVEL OF SWAPPING	DEPENDS ON LEVEL OF SWAPPING	DEPENDS ON LEVEL OF IMPUTATION
Log-linear model	NO OVERALL PATTERN	NO OVERALL PATTERN	NO OVERALL PATTERN

4.3 Impact on Origin-Destination Tables

Origin destination tables are different to census area statistics tables in that they consist of data for all combinations of areas in England and Wales (in each O/D table), and can have over 10 million cells depending on the breakdown of variables and level of geography, they are extremely sparse. Zeros typically comprise 98-99% of the table cells at OA level with small cell values making up the majority of non-zero cells though zero rows are usually suppressed from output (summary statistics are shown in Appendix A1.1). Appendix A3.3 provides a description of O/D tables.

Determining an appropriate SDC strategy for O/D tables is very problematic due to their sparsity. In the final results section of this report, we examine what impact the methods of over-imputation and record swapping have on this type of output. For simplicity, in this interim report, we concentrate on the total flows between origin and destination because these are sufficient to illustrate the main differences between the SDC methods. By total flows, it is meant the total numbers of flows between origin and travel-to-work destination (and not the variable breakdown, e.g. breakdown of flows into numbers travelling by bike, bus, etc).

Disclosure Risk

Group, negative attribute and within-group disclosure assess risk arising from rows or columns where the majority of cells are zero. Since there are millions of cells in the O/D tables which are zero, these risk measures are not appropriate here and instead we focus on the percentage of cells unperturbed that were not originally zero, and the percentage of ones unperturbed. The latter corresponds to the principle of identity disclosure. In summary;

- Random swapping with SCA results in only 1.3% of total flows being **un**perturbed. This is because many of the total flows are either ones or twos which are small cell adjusted. This is a key problem with SCA as flows are actually 'disappearing'.

- Over-imputation results in 67-70% of the totals flows being unperturbed for a 20% imputation. Because geography is an imputed variable, the locations of the households are deleted (origins) and new locations imputed based on the remaining data. The work locations are unchanged (destinations). Thus records where geography is imputed may result in new flows being created. For example;

Before imputation

Person 1	Married	Age 42	Lives in location X	Male	Travels by bike	Works in location A
Person 2	Single	Age 21	Lives in location Y	Female	Travels by bus	Works in location B

Imputation of geography for two records

Person 1	Married	Age 42	**Lives in location Z**	Male	Travels by bike	**Works in location A**
Person 2	Single	Age 21	**Lives in location W**	Female	Travels by bus	**Works in location B**

Geography is imputed for both persons 1 and 2 which results in new flows being created.

- Swapping of geography (**i.e. picking up one household and putting it in the location of another and vice versa**) results in the total flows being completely unchanged. This is because the locations of the characteristics of the households are swapped (origins) but the work locations are unswapped (destinations). Thus all flows still remain intact but the characteristics of the households making those flows are changed; e.g. whether a flow from X to A involves travel by bike or bus. For example;

Before swapping

Person 1	Married	Age 42	Lives in location X	Male	Travels by bike	Works in location A
Person 2	Single	Age 21	Lives in location Y	Female	Travels by bus	Works in location B

Swapping of geography for two records

Person 2	Single	Age 21	**Lives in location X**	Female	Travels by bus	**Works in location A**
Person 1	Married	Age 42	**Lives in location Y**	Male	Travels by bike	**Works in location B**

Persons 1 and 2 have swapped location but the flows between origin and work location remain.

Data Utility

O/D tables are extremely sparse so many of the Infoloss software measures of utility would not be appropriate being heavily influenced by the extreme proportion of zeros. Instead we only examine the frequency distribution of the absolute differences between the original and protected cell values. Appendix 3.3 presents some results. In summary:

- Imputation results in 73-91% of the perturbed cells (total flows only) having an absolute difference of one.
- Random swapping with SCA results in 57% of the perturbed cells (total flows only) having an absolute difference of one but 39% of the perturbed cells having an absolute difference of two or three. This is due to SCA modifying cells by -/+1 or -/+2 on top of swapping. Therefore this approach results in the greatest distortion.
- Swapping alone results in no change to the perturbed cells (total flows only) as described above for disclosure risk.

Conclusions

- Record swapping alone does not produce enough protection to O/D tables as the total flows are unchanged and thus a count of one in the total flow column may potentially lead to disclosure.
- Record swapping with SCA severely distorts the very sparse tables with many absolute differences of two or more on the small cell counts, of which comprise the majority of the non-zero part of the table.
- Both swapping and imputation have severe effects on data utility in that these methods are likely to create inconsistencies in the O/D tables, for example a flow being created that is nonsensical e.g. a student travelling by foot a very long distance to a work location that is not a university/school/college[4]. We note that in 2001 England and Wales O-D tables were based on 'travel to work' whereas Scotland based theirs on 'travel to work or study'.
- Imputation distorts both total flows and the breakdown of these flows by variable (e.g. by method of travel to work) thus creating inconsistencies in both the location of flows and inconsistencies to the breakdown of the type of flows. Thus the distortion is worse for imputation than swapping.

5. Discussion, Recommendations and Further Work

5.1 Discussion

- In terms of disclosure risk, over-imputation has comparable results with record swapping for tables 1 and 2. Table 3 which included age showed a more pronounced reduction in disclosure risk for imputation compared to swapping (since age was imputed as well as geography) but likewise the impact on distortion to utility was more pronounced. Thus imputing more variables is another way (other than increasing the perturbation percentage for example) to achieve an appropriate balance between risk and utility.
- Targeted imputation in particular appeared to work particularly well in reducing the risk of identity disclosure.
- Swapping and over-imputation have similar effects in that generally they homogenise the data. In terms of swapping this is because households are essentially shuffled around within LADs so the statistical effect is of homogenising to the LAD mean. In terms of imputation, blanked values for geography are imputed using a donor from the same LAD. However the homogenising effect tends to be far greater for imputation as evidenced by the results on tables 1 to 3. This is because these blanked values are replaced using the remaining data (in effect replicating existing values), which for high levels of imputation has a large impact on utility.
- Record swapping using control strata preserves statistics relating to these strata which is an important advantage (compared to imputation) for users who require certain key statistics.

[4] A further example of a seemingly nonsensical combination of origin, destination and mode of travel – is where someone lives in Sheffield, travels down to Southampton at the weekend, walks to work each day in Southampton so the O-D matrix would show origin = Sheffield, destination = Southampton and walks to work.

- Over-imputation is not a well-known approach and has not previously been tested on UK census data. Swapping is a tried-and-tested approach both in the UK and in international NSIs.

- We refer to a further disadvantage of imputation compared to swapping which is the potential for edit failures since both age and geography were imputed.

- Small cell adjustment in combination with record swapping has several advantages, not only is this approach partially transparent to users (small cell counts removed from the tables), it also negates some of the effects of swapping on utility. In addition, SCA can be controlled with respect to preserving totals if required. However SCA results in a loss of consistency between totals in different tables which is strongly disliked by users.

- Both record swapping (with or without SCA) and over-imputation are unsuitable for O/D tables as inconsistencies are likely to be created in the form of nonsensical flows. In addition, both record swapping with SCA and over-imputation result in much distortion, in terms of absolute differences, to the small cell counts.

- In conclusion, although over-imputation provides another parameter to allow reduction of disclosure risk through imputation of more variables, the effects on data utility are substantial. Statistics cannot be preserved through control strata, there is the potential for edit failures, and most importantly data are 'lost' as they are being deleted and replicated from the remaining records. In contrast, record swapping introduces uncertainty into the geography-attribute relationships but this can be limited using controls. In addition, the attribute data alone are unaffected and distortion to variance and association is minimal compared to imputation. Moreover totals and sub-totals are approximately or entirely preserved with swapping.

5.2 Recommendations

- **We recommend over-imputation be dropped from the short-list based on the significant disadvantages in terms of data utility illustrated in tables 1, 2 and 3. The reduction in disclosure risk via imputation of more variables can be achieved by swapping a larger percentage of records. Dropping over-imputation from the short-list would free up more time to concentrate on whether record swapping or the ABS cell perturbation method is most appropriate for disclosure control of the outputs from the 2011 Census.**

- We recommend an alternative approach be considered for O/D tables such as releasing under a licence or access agreement. O/D tables are very different to other census outputs in that they are extremely sparse and the small cell counts are likely to be severely distorted by any kind of SDC method as illustrated by the analysis in this report.

- Given that over-imputation and record swapping at the 20% and 10% levels do not reduce disclosure risk significantly (for example the percentage of ones that are still ones is always still above 50%), it is likely that a targeted approach would be recommended rather than random.

5.3 Suggestions for Further Work

- In light of the advantages of small cell adjustment in combination with record swapping, along with the disadvantage of loss of consistency, a further approach could be considered. This possibility is to use ABS cell perturbation with swapping to achieve the same effect as swapping with small cell adjustment but resulting in consistency across tables. This would obviously depend on the risk-utility results as yet to be evaluated.

- In terms of other future work, we plan to continue the quantitative evaluation looking at a variety of other census tables as proposed in the Miller et al. (2007) report. The tables need to be prioritised in terms of order of importance for user analysis.

- Moreover an assessment needs to be undertaken of the protection provided by the SDC methods from disclosure by differencing.

- A further report is planned with updated results on risk-utility in August 2008.

Appendix

A1.1 Summary Statistics Describing the Four Census Tables

	Table 1 – ORCD data	Table 1 – CPCD data	Table 2 – ORCD data	Table 2 – CPCD data
Total number of cells	2,240	1,760	17,844	95,168
Small cells	12%	12%	10%	3%
Zeros	20%	8%	63%	58%
Average cell size	364	463	10	37
Standard error (average cell size)	22.82	30.22	0.17	0.78

	Table 3 – ORCD data	Table 3 – CPCD data	Table 4 – ORCD data	Table 4 – CPCD data
Total number of cells	95,168	57,792	10,739,114 (total flows only)	6,521,466 (total flows only)
Small cells	22%	16%		
Zeros	24%	20%	99%	99%
Average cell size	5	7		
Standard error (average cell size)	0.02	0.04		

A2.1 Full Methodology for Over-Imputation

A new method of over-imputation had to be devised for disclosure control of census data as it is an unknown approach. Imputation in general is a very complex procedure, one reason being the relationships that exist between variables. For example, imputing ethnicity is not straightforward as this could potentially lead to errors for country of birth, and other correlated variables.

Moreover CANCEIS is designed to impute the 'best' possible values based on a nearest neighbour donor as close as possible to the true values. Thus for categorical values such as ethnicity, housing type (detached, semi-, terraced, flat…), etc, it is likely that the exact value will be imputed – providing no protection. The variables *age* and *geography* were chosen for imputation because there are a wide range of values along the scale that are possible e.g. age of 60 might be imputed with a value of 57,58,59,60,61,62,63 rather than a few very different choices with ethnicity or housing type e.g. a housing type detached might be imputed as flat or terraced which may not be plausible. It is likely a value close to the original will be imputed, giving some protection but not distorting the data too severely. In addition, *geography* is commonly associated with disclosure risk as at low levels, it can be used to help identify individual households and persons.

Over-imputation was carried out six times for the SJ EA. 2%, 10% and 20% random samples of households were selected within strata of LAD and number of persons in household. These strata were used in order that over-imputation had some degree of comparability with record swapping (since record swapping will be based on swapping households within LAD and secondly in each strata (LAD by size of household) all households (and hence all persons) had an equal probability of selection. Over-imputation was then repeated using the population of high risk households[5] (targeted imputation). The smaller samples were drawn from the larger 20% samples, to avoid introducing variance between them. The methodology is as follows:

Step 1: Blank out the values of the variables *age* (and *year of birth*) and all geography variables except *census district code*, *district code* and *county code*, for the sample of households in the strata only.

Step 2: For each sampled household, one at a time, impute *age* (and therefore *year of birth*) based on all remaining variables for the household except geography. N.B. geography is not used here, so that a wider population is used to find donors for the missing ages.

Step 3: Impute *ed code* (ED) and *ward* for the sampled households (one household at a time) based on match variables of imputed ages, existing *census district code*, *district code* and *county code* and all other household variables.

[5] High risk records are defined in the ORCD file as those which make up the small cell counts in tables of religion/age/sex/OA, travel to work/age/sex/OA, country of birth/sex/OA, economic activity/sex/llti/OA, health status/age/sex/OA. The *address* variable is used to match the high risk records in the ORCD file to the CPCD file.

The targeted imputation follows the same procedure but using the sample of risky households instead.

In summary, after over-imputation was applied to the CPCD file, households which were selected had *ed code* and *ward* imputed but they remained within the same LAD. After over-imputation households which were selected had age imputed: approximately 10% of these ages had exactly the same value imputed back (no change), approximately 45% had an age within one to four years difference from the original imputed, 30% had an age five to ten years difference from the original imputed, the remaining approximately 15% had an age greater than 10 years different from the original value imputed. CANCEIS aims to minimise the possibility of edit failures (e.g. a 10-year old child married to an 80-year old adult).

A2.2 Full Methodology for Record Swapping

Record swapping had been previously carried out by a consultant from Southampton University (Natalie Shlomo). A random sample within strata defined by control variables was selected using a fixed swapping rate *f*. The control variables that were used were: hard-to-count index[6], household size, sex and broad age distribution of the household (0-25, 25-44, 45 and over). For each household selected, a paired household is found. The effect of using strata is that households are paired matching on the four control variables.

Then all geographical variables in all selected records were swapped – i.e. **all geography variables related to the location of the household** (address, OA, LAD, etc). This has the same effect as swapping all other variables and leaving geography fixed. The following percentage of records were swapped in total: 2%, 10% and 20%. This meant that samples of 1%, 5% and 10% had to be found and paired with another 1%, 5% and 10% (respectively). As with over-imputation, the smaller samples were sub sampled from the largest sample in order to avoid introducing variance between them.

For the targeted swap, based on a set of standard census tables (see footnote 4): small cells in the tables were identified and flagged. A targeted record swap was implemented by pairing and swapping households that matched not only on the control variables but also on the flagged variable. If, however, a household that was selected for swapping did not have a match on the control variables from among the flagged households, a match was found outside the flagged households. Please note that the targeted over-imputation method was also based on the same population of high risk records.

A3.1 Disclosure Risk Measures

The disclosure risk measures are based on comparing the original (raw) tables with the protected tables.

(i) Identity disclosure: picks up single respondents in the table that are single respondents in the same cell in the protected table.

(ii) Group disclosure: picks up cases where all respondents fall in one cell in a row (or column) and the same pattern in the same cells in the protected table.

(iii) Negative attribute disclosure: picks up rows (or columns) which contain only zeros and similarly in the protected table.

(iv) Probability of change: Number of original cells of 1 or 2 that keep the exact same value in the perturbed table divided by the total number of original cells of size 1 or 2.

(v) *Within-group disclosure: picks up rows (or columns) where there is a single respondent in a cell with all other respondents falling in another cell, and the same occurring in the protected table (not yet evaluated).*

Measure (i) is presented as a percentage, dividing by the number of cell counts of one. Measure (ii) is expressed as the number of cases of (ii) dividing by the number of 'risky' rows (those which present potential disclosure in the original table) to give relative values.

Results have not been included for the 2% perturbation levels. Moreover there were no risky rows in the original table in the case of (ii) – group disclosure and (iii) – negative attribute disclosure, for table 1, and furthermore no risky rows in the original table in the case of (iii) for table 2. Thus no results are shown. It is hoped that the remaining tables to be analysed (as proposed in the Miller et al. (2007) report) will show more cases of (ii) and (iii) although it is important to note that incidence of these types of disclosure is generally quite unusual.

Disclosure Risk Analysis on Table 1

[6] The hard-to-count index was constructed from census variables known to be associated with under-enumeration.

	Number of cases of (i)	Probability of (iv)
Random swapping 20%	79%	0.72
Random swapping 10%	90%	0.84
Random swapping 20% with SCA	0%	0
Random swapping 10% with SCA	0%	0
Targeted swapping 20%	81%	0.68
Targeted swapping 10%	89%	0.81
Random imputation 20%	84%	0.80
Random imputation 10%	91%	0.90
Targeted imputation 20%		0.51
Targeted imputation 10%	76%	0.70

Disclosure Risk Analysis on Table 2

	Number of cases of (i)	Number of cases of (ii)	Probability of (iv)
Random swapping 20%	71%		0.77
Random swapping 10%	81%	13 / 13 rows	0.85
Random swapping 20% with SCA	0%		0
Random swapping 10% with SCA	0%	5 / 13 rows	0
Targeted swapping 20%	72%	11 / 13 rows	0.74
Targeted swapping 10%	83%	11 / 13 rows	0.84
Random imputation 20%	68%	6 / 6 rows	0.75
Random imputation 10%	81%	6 / 6 rows	0.88
Targeted	72%	3 / 6 rows	0.70
Targeted imputation 10%	82%	5 / 6 rows	0.87

Disclosure Risk Analysis on Table 3

	Number of cases of (i)	Number of cases of (ii)	Number of cases of (iii)	Probability of (iv)
Random swapping 20%	76%	1 / 1 rows	1 / 1 rows	0.72
Random swapping 10%	88%	1 / 1 rows	1 / 1 rows	0.85

Random swapping 20% with SCA	0%	0 / 1 rows	1 / 1 rows	0
Random swapping 10% with SCA	0%	0 / 1 rows	1 / 1 rows	0
Targeted swapping 20%	78%		1 / 1 rows	0.72
Targeted swapping 10%	88%	1 / 1 rows	1 / 1 rows	0.84
Random imputation 20%	66%	0 / 0 rows	0 / 0 rows	0.58
Random imputation 10%	80%	0 / 0 rows	0 / 0 rows	0.74
Targeted imputation 20%	58%	0 / 0 rows	0 / 0 rows	0.52
Targeted imputation 10%	74%	0 / 0 rows	0 / 0 rows	0.69

N.B. Disclosure risk results for table 4 (the O/D table) are contained in Appendix section A3.3 along with a full explanation of this type of table and the data utility results.

A3.2 Results for Data Utility Analysis on Short-listed Methods
Data Utility Impact on Table 1

	Association Measure		(Cell) Distance Metrics				Changes to subtotals			
	% relative difference in Cramer's V -(orig-prot)	Average of ratios of variance across geographies (prot/orig)	Absolute Average Deviation	Hellingers' distance	Relative Absolute Deviation	Rows (geog'hies) which have changed rank group	Absolute difference: sex	Absolute difference: religion	Absolute difference: country of birth	Ratio of deviance (log-linear model all var but geog): prot/orig
Random swapping 2%	-1.01%	0.9990	0.4431	0.4439	3.4830	107	0	0	0	1.0001
Random Swapping 10%	-1.32%	0.9923	2.395	1.7589	16.0250	377	0	0	0	1.0034
Random Swapping 20%	-1.85%	0.9791	4.5113	3.3453	31.9000	656	0	0	0	1.0133
Random swapping 2% with SCA	1.90%	1.0653	3.9955	1.2057	3.6268	158	16	8.25	16	1.0014
Random swapping 10% with SCA	2.11%	1.0044	5.6537	2.2091	16.1536	391	13	4	13	1.0458
Random swapping 20% with SCA	2.37%	1.0443	7.2747	3.6623	32.0103	662	4.50	6.37	17.50	1.0144
Targeted Swapping 2%	-0.79%	0.9945	0.5506	0.4713	3.6411	110	0	0	0	0.9999
Targeted Swapping 10%	-1.32%	0.9860	2.7705	1.8834	16.5196	433	0	0	0	1.0037
Targeted Swapping 20%	-1.58%	0.9850	4.5921	3.4243	32.4429	695	0	0	0	1.0164
Random Imputation 2%	-9.14%	0.9158	2.0662	2.6050	28.7193	263	23824	5956	23824	0.9239
Random Imputation 10%	-9.41%	0.9137	2.8363	2.7407	29.1636	354	23890	5973	23890	0.9235
Random Imputation 20%	-11.83%	0.9141	4.1086	3.0402	30.0057	492	24043	6011	24043	0.9228
Targeted Imputation 2%	-7.25%	0.9175	2.2170	2.6119	28.7523	312	23763	5941	23763	0.923
Targeted Imputation 10%	-8.87%	0.9143	4.6911	3.1110	29.6898	531	23874	5969	23874	0.9233
Targeted Imputation 20%	-11.29%	0.9100	7.3814	3.7140	31.5193	756	24128	6032	24128	0.9226

Data Utility Impact on Table 2

	Association		(Cell) Distance Metrics				Changes to subtotals		
	% relative difference in Cramer's V -(orig – prot)	Average of ratios of variance across geographies (prot/orig)	Absolute Average Deviation	Relative Absolute Deviation	Hellingers' distance	Rows (geog'hies) which have changed rank group	Absolute difference: persons in household	Absolute difference: acc'dation type	Ratio of deviance (log-linear model all var but geog): prot/orig
Random swapping 2%	-0.06%	0.9976	0.1677	0.1135	0.1371	632	0.66	0.50	0.9885
Random Swapping 10%	-0.09%	0.9913	0.6305	0.4428	0.4139	2138	2.66	3.50	0.9528
Random Swapping 20%	-0.10%	0.9876	1.1723	0.8192	0.6952	3589	5.33	12.00	0.9225
Random swapping 2% with SCA	0.75%	1.0248	0.2936	1.4086	0.5993	967	16.33	19.75	1.0224
Random swapping 10% with SCA	1.89%	1.0033	0.7446	1.6268	0.7312	2116	33.00	19.75	0.9853
Random swapping 20% with SCA	3.61%	1.0103	1.2755	1.9105	0.9238	3221	37.33	28.00	0.9557
Targeted Swapping 2%	-0.02%	0.9966	0.1618	0.1015	0.1313	555	0.66	1.00	0.9876
Targeted Swapping 10%	-0.07%	0.9898	0.6293	0.4194	0.4098	2049	2.66	2.50	0.9520
Targeted Swapping 20%	-0.10%	0.9873	1.1751	0.7824	0.6847	3394	2.00	2.50	0.9203
Random Imputation 2%	-3.08%	0.9980	0.7508	0.1482	0.2975	454	309	232	1.0011
Random Imputation 10%	-3.31%	0.9879	1.8293	0.4543	0.6766	1263	938	703.50	0.9997
Random Imputation 20%	-3.42%	0.9808	2.8282	0.7949	1.0477	2108	1670.66	1253	1.0003
Targeted Imputation 2%	-0.24%	0.9974	0.7788	0.1763	0.3181	519	281.66	211.25	1.0010
Targeted Imputation 10%	-0.70%	0.9910	2.0098	0.5436	0.7920	1485	1002.66	752	1.0058
Targeted Imputation 20%	-1.39%	0.9743	3.1879	0.9416	1.2324	2490	1694.5	2259	1.0100

Data Utility Impact on Table 3

	Association		(Cell) Distance Metrics				Changes to subtotals			
	% relative difference in Cramer's V -(orig – prot)	Average of ratios of variance across geographies (prot/orig)	Absolute Average Deviation	Relative Absolute Deviation	Hellingers' distance	Rows (geog'hies) which have changed rank group	Absolute difference: age	Absolute difference: sex	Absolute difference: marital status	Ratio of deviance (log-linear model all var but geog): prot/orig
Random swapping 2%	0%	1.0017	0.0898	1.016	0.3762	4285	3.62	4	6	1.0016
Random Swapping 10%	-0.8%	1.0084	0.3681	4.4977	1.0021	17807	6.12	10	33	1.0089
Random Swapping 20%	-1.9%	1.0218	0.6563	8.2739	1.5005	30606	17	21	52	1.0273
Random swapping 2% with SCA	0%	1.0284	0.3793	15.2119	2.3653	12814	118.05	118.50	48.50	1.1684
Random swapping 10% with SCA	0.7%	1.0583	0.6311	17.5193	2.5539	23771	36.66	34	163	1.1758
Random swapping 20% with SCA	1.3%	1.0464	0.8878	19.8739	2.7855	33794	136.50	136.50	34.56	1.193
Targeted Swapping 2%	-0.7%	1.0009	0.0891	1.0190	0.3793	4371	3.5	2	10	1.0009
Targeted Swapping 10%	-3.0%	1.0080	0.3646	4.5559	0.9962	17977	8.11	6	28	1.0066
Targeted Swapping 20%	-4.70%	1.0194	0.6534	8.3390	1.5025	30588	13.33	9	30	1.0243
Random Imputation 2%	-3.79%	0.9945	0.2407	2.0549	0.5988	5748	468	468	60	1.0023
Random Imputation 10%	-4.08%	0.9856	0.7086	6.9279	1.3251	17591	1411	1411	192	1.0152
Random Imputation 20%	-4.09%	0.9813	1.0918	11.1099	1.8993	25189	2514	2514	350	1.0327
Targeted Imputation 2%	-4.00%	0.9971	0.2522	2.2744	0.6370	6273	423	423	63	1.0068
Targeted Imputation 10%	-4.07%	0.9949	0.7682	7.8629	1.4813	19172	1519	1519	246	1.0364
Targeted Imputation 20%	-4.15%	0.9857	1.1863	12.5525	2.1324	26595	3402	3402	525	1.0802

General Observations from Results on Data Utility

Measures of Association

- Both swapping and imputation reduce the level of association; this demonstrates the general flattening of proportions towards the average (proportions refers to the cross-classifications of the table e.g. non-UK Christian females) which means the variables in the table are becoming more independent.
- Imputation reduces this association to a much greater degree – particularly random imputation.
- SCA in combination with record swapping appears to result in an increase in the level of association suggesting the two methods have opposing effects.

Distance Metrics

- Targeted methods result in greater distortion to the distance metrics in table 1. In table 3, targeted imputation produces greater distortion than random imputation but considering the tables together, there appears to be no overall pattern. Although it is important to remember these metrics are averaged so can mask individual cell variations.
- Small cell adjustment in combination with record swapping increases the distortion to the distance metrics, compared to swapping applied alone.
- Imputation generally results in a greater magnitude of distortion than swapping, likely because there is the potential for the blanked values to be replaced with a donor that is very different, unlike swapping where paired, matched households are found.

Change in Variance

- Both swapping and imputation reduce variance (although it slightly increases for table 3 and swapping), i.e. distributions of counts in the rows are flattening out.
- The variance, in general, decreases more with imputation than it does with swapping. This is likely because the existing data are used to replace blanked values with imputation and thus they become much more homogeneous whereas swapping only switches geographical location.
- Targeted methods reduce the variance to a greater degree than a random approach, possibly because they concentrate on one end of the distribution.
- SCA counteracts this effect so that variance increases.

Impact on Rankings by Geography

- There appears to be little difference between the impact of the targeted swapping as compared to random swapping. In the case of imputation, the targeted approach does much worse, with more geographies moving between rank groups.
- There appears to be no noticeable difference between imputation and swapping in terms of impact on rankings, possibly because they both involve perturbing geography and at the same level of geography (between small areas within LADs).

Changes to Subtotals

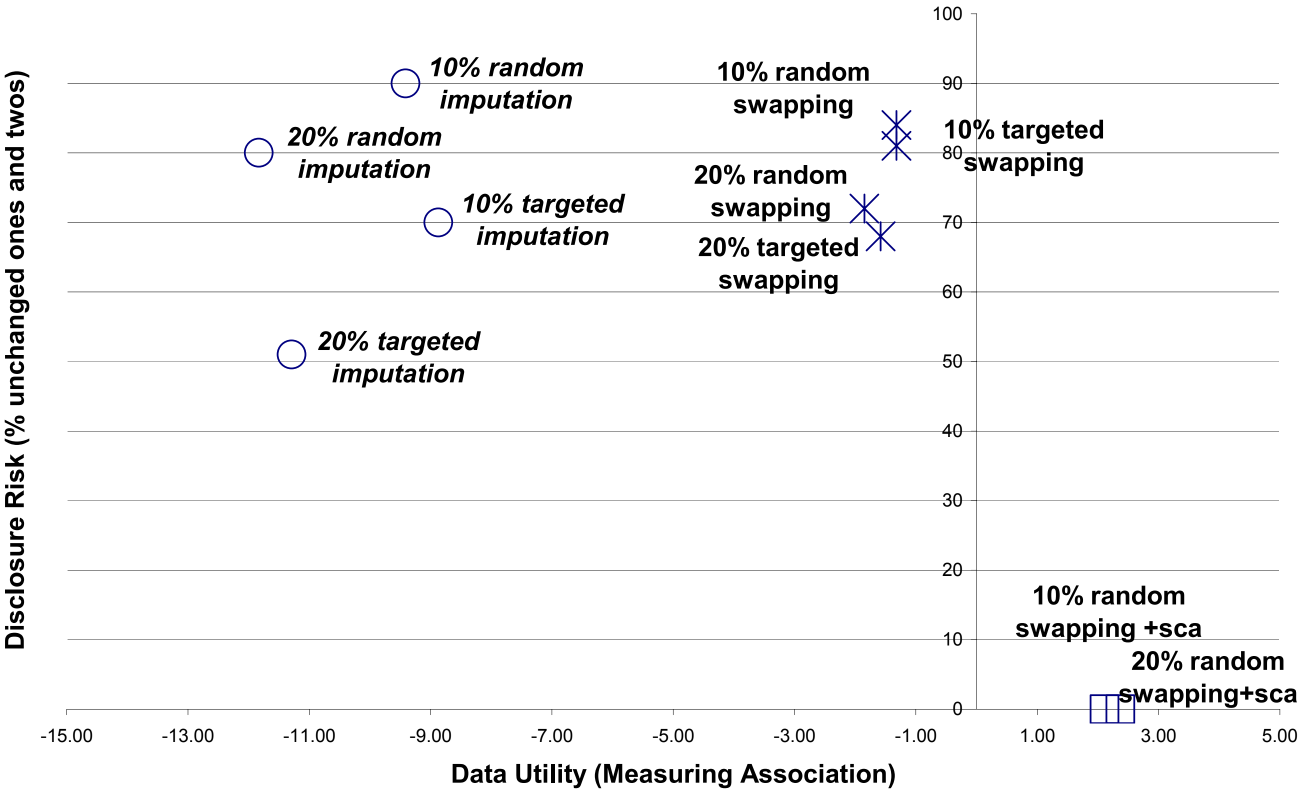
- The tables show clear results with small changes in the case of swapping, if any, to the subtotals of the variables. This is because households are only moved around geographically with swapping, so the average differences at higher levels should be approximately zero. As would be expected, there is some change with SCA.
- In contrast, there are very large changes with imputation as people are being added into areas that weren't there before. This is particularly apparent in table 1.

Changes to Log-linear Model

- There is no clear pattern in terms of the impact of swapping and imputation on the ratio of deviance. However the impact in table 3 is greater for imputation than swapping, as would be expected since age was imputed.

The differences between the SDC methods can be observed in a Risk-Utility map – figure 1, relating to table 1 (similar results can be seen for the other tables). Disclosure Risk is measured in terms of the percentage of ones and twos that are unchanged between the original and protected tables (risk measure iv). Data Utility is measured in terms of the percentage relative difference in measure of association (using Cramer's V), which captures the homogenising effect of both imputation and swapping.

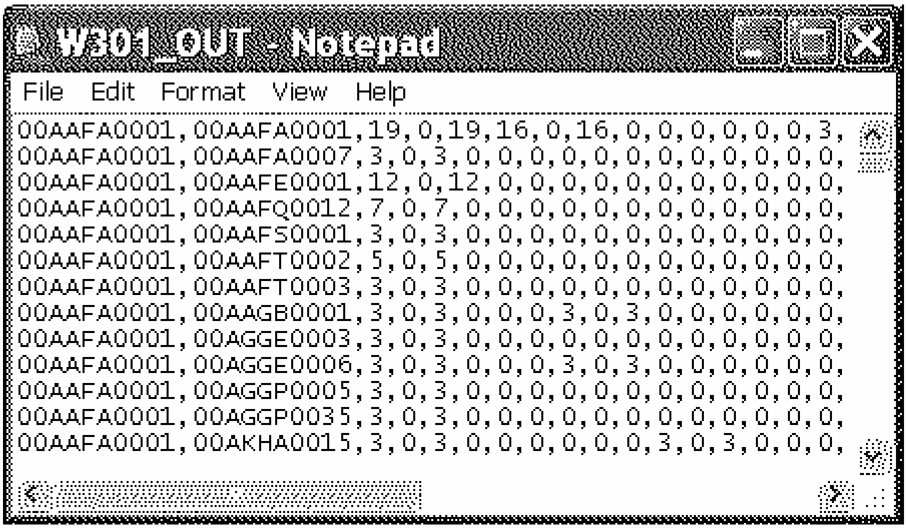
Figure 1: Disclosure Risk – Data Utility Map for 20% and 10% level of Record Swapping and Over-Imputation.



The map shows clearly how the disclosure risk is reduced to zero when swapping is combined with small cell adjustment. Comparing swapping on its own with imputation, there is not much difference between levels of disclosure risk, although the targeted methods are better at reducing risk than the random approaches. In terms of utility, the imputation methods are clearly worse, with the association being changed significantly from zero whereas swapping, and swapping with small cell adjustment have data utility close to zero (only a slight change in the level of association).

A3.3 Risk-Utility Results for Table 4 (Origin-Destination: Total Flows Only)

Origin Destination tables typically have a format as shown below where the first column is the origin (home) and the second column is the destination (travel to work location). Flows that are zero (or adjusted to zero) are suppressed. The remaining columns relate to cells 1,2,3,4, etc in the table layouts below, for example, the migrants O/D table has 12 columns altogether. Thus column 8 would relate to the flows between OAs for males aged 16 to pensionable age.



```
00AAFA0001,00AAFA0001,19,0,19,16,0,16,0,0,0,0,0,0,3,
00AAFA0001,00AAFA0007,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AAFE0001,12,0,12,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AAFQ0012,7,0,7,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AAFS0001,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AAFT0002,5,0,5,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AAFT0003,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AAGB0001,3,0,3,0,0,0,3,0,3,0,0,0,0,0,0,
00AAFA0001,00AGGE0003,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AGGE0006,3,0,3,0,0,0,3,0,3,0,0,0,0,0,0,
00AAFA0001,00AGGP0005,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AGGP0035,3,0,3,0,0,0,0,0,0,0,0,0,0,0,0,
00AAFA0001,00AKHA0015,3,0,3,0,0,0,0,0,0,3,0,3,0,0,0,
```

Migrants - Age and Sex

	All people	Male	Female
Total	1	2	3
0-15	4	5	6
16-pensionable age	7	8	9
Pensionable age and above	10	11	12

The risk-utility results below are shown in terms of the percentage of cells left unperturbed. The 20% swaps and imputations are shown only to illustrate the effects for the higher levels of perturbation.

Risk-Utility Results for the O/D table.

	% cells unperturbed (that were not originally zero)	% ones unperturbed
Random swapping 20%	100%	100%
Random swapping 20% with SCA	1.3%	0%
Targeted Swapping 20%	100%	100%
Random Imputation 20%	70.1%	80%
Targeted Imputation 20%	67%	77%

Impact due to Swapping

- In terms of totals (column 1 of the O/D table), swapping leaves these unaffected i.e. the percentage of cells unperturbed is 100%, because although the households are swapped, the flow is still there (just the swapped households have different characteristics).
- In terms of variable breakdowns (which have not yet been analysed), the extent to which these are damaged is likely to be similar as for previous tables where swapping has been applied.
- Swapping will add uncertainty to the variable breakdowns so an internal flow of 1 may not be a true flow of 1.
- Swapping is likely to lead to inconsistencies in the flows (where control variables are not relevant); for example a person travelling by bike a very long distance or a student travelling to an area where there isn't a college or university [see comment in the Conclusions to Section 4].

Impact due to Swapping with SCA

- In terms of total flows, SCA with swapping results in the disclosure risk being reduced to a minimal level with the percentage of cells unperturbed being 1.3%.
- As expected the percentage of ones being unperturbed is zero.

Impact due to Imputation

- In terms of totals, imputation (in this case) removes geography relating to the origin and thus a flow disappears, and is replaced by a new origin so a new flow is created.
- In terms of variable breakdowns, the extent to which these are damaged is likely to be similar as for previous tables where imputation has been applied.
- Imputation adds uncertainty to the variable breakdowns so a flow of 1 may have been imputed.
- Imputation is likely to lead to inconsistencies in the flows as with swapping.

In summary, the most disclosive scenario may be represented by a one in the total flows column. Swapping does not change this one but there may be uncertainty as to where the true flow lies in terms of variable breakdown. Imputation may remove ones in the total flows (or add them in) and thus provides more protection in this respect but also more damage. Swapping with small cell adjustment removes some ones altogether so may be thought to offer the most protection since the risk is effectively reduced to zero.

Frequency Distribution – Cell Differences

O/D tables are extremely sparse so many of the Infoloss software measures of utility would not be appropriate. Instead we only examine the frequency distribution of the absolute differences between the original and protected cell values. These are shown for swapping with SCA, and imputation only, since swapping alone has no impact on the total flows. Results are illustrated for the 20% imputation rate only, where the impact is greatest.

	Percentage of cells no change	Absolute difference = 1	Absolute difference = 2	Absolute difference = 3	Absolute difference = 4	Absolute difference = 5	Absolute difference = 6+
Targeted 20% imputation	99.5%						
Of cells that changed value		91%	5%	2%	1%	1%	0%
Random 20% Imputation	99.6%						
Of cells that changed value		73%	14%	5%	2%	1%	5%
Random Swapping 20% with SCA	98.4%						
Of cells that changed value		47%	10%	29%	3%	1%	10%

There are so many zeros in the table that the proportion of non-zeros cells that change are very small in comparison. Targeted imputation results in many more absolute differences of larger magnitudes and swapping with SCA to an even greater extent.

References

1. 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.

2. Miller, C. Longhurst, J. Tromans, N. and Young, C. (2007) Quantitative Risk-Utility Evaluation of SDC Methods for 2011 Census – report for UK CDMAC

Acknowledgements

Many thanks to the Edit and Imputation branch at ONS who carried out the over-imputation work for this analysis.