



Evaluating a statistical disclosure control (SDC) strategy for 2011 Census outputs

Evaluating a statistical disclosure control (SDC) strategy for 2011 Census tabular outputs

Table of Contents

Executive summary

Summary

- 1 Background**
- 2 SDC methods**
 - 2.1 Record swapping**
 - 2.2 Imputation**
 - 2.3 ABS cell perturbation**
 - 2.4 Small cell adjustment**
- 3 Approach**
 - 3.1 Evaluating risk and utility quantitatively**
 - 3.2 Evaluation**
- 4 Summary of results**
 - 4.1 Mandatory evaluation criteria**
 - 4.2 Secondary evaluation criteria**
 - 4.3 Origin-destination tables**
 - 4.4 Other considerations**
 - 4.5 Evaluation summary**
- 5 Future work**

Appendix A Methods

- A1 Record swapping**
- A2 Over-imputation**
- A3 ABS cell perturbation**

Appendix B Quantitative evaluation

- B1 The data**
- B2 Disclosure risk measures**
- B3 Utility measures**

Appendix C Results

- C1 Risk**
- C2 Utility**
- C3 Disclosure risk and utility**
- C4 Origin-destination tables**

Appendix D Assessment criteria with scoring

Appendix E Glossary and abbreviations

Evaluating a statistical disclosure control (SDC) strategy for 2011 Census tabular outputs

Executive summary

(1) Introduction

This paper provides an evaluation of three possible methods of Statistical Disclosure Control (SDC) to be applied to outputs from the 2011 Census. The Registrars General (RsG) of England and Wales, Scotland, and Northern Ireland had expressed the aim to have the same SDC method applied to the 2011 Census in each country of the UK. The evaluation resulted in the recommendation to use record swapping as the primary method, and this has been agreed to by the RsG.

(2) Background

Users were critical of the lack of harmonisation in the 2001 Census, when ONS and NISRA (Northern Ireland Statistical and Research Agency) made a late decision to apply small cell adjustment (SCA) on top of record swapping, while using higher geographic thresholds than GROS (General Register Office for Scotland), who did not apply SCA. As a result problems were caused for users of 2001 Census products.

A work-package was established to address these issues. In November 2006 the RsG agreed to aim for a common UK SDC methodology for 2011 Census outputs. The RsG considered that, as long as there has been systematic perturbation of the data, the guarantee in the Code of Practice would be met. It was therefore agreed that small counts (0's, 1's, and 2's) could be included in publicly disseminated census tables provided that

a) uncertainty as to whether the small cell is a true value has been systematically created; and

b) creating that uncertainty does not significantly damage the data.

Though pre- and post-tabular methods could be considered, the RsG expressed a preference for pre-tabular methods, provided there is not undue damage to the data, and reported the preference for SDC to have a 'light touch'.

Since agreeing this UK SDC Policy, the Statistics and Registration Service Act (SRSA) 2007 has come into force. Section 39 (2) of the Act defines personal information as information which relates to and identifies a particular person. It specifies what constitutes a disclosure of information and the sanctions that may apply for any breach of confidentiality. Following the introduction of the SRSA, the National Statistics Code of Practice has been superseded by the Code of Practice (CoP) for Official Statistics (January 2009), which affirms as one of its eight fundamental principles that 'Private information about individual persons (including bodies corporate) compiled in the production of official statistics is confidential, and should be used for statistical purposes only'. The UK SDC Policy is in line with both Section 39 of the SRSA and the

Code of Practice.

Throughout the evaluation exercise, ONS ensured that there was consultation with the other UK Census Offices by setting up the UK SDC Working Group, which includes representatives from GROS, NISRA and Welsh Assembly Government (WAG), and with the UK Census Design and Methodology Advisory Committee SDC sub-group (UKCDMAC), which consists of statisticians and academics. Input from both groups was incorporated into the evaluation.

(3) Overview of evaluation

The three methods evaluated were record swapping, over-imputation and a post-tabular method of cell perturbation. This last was a modified version of the method developed by the Australian Bureau of Statistics (ABS). The modification improved the utility of the method. See sections 2.3 and A3 for details. It is referred to as IACP (invariant ABS cell perturbation). These three methods were short-listed from a wide-range of SDC methods, the short-list having been agreed with the RsG.

For record swapping and over-imputation, both random and targeted techniques were examined. Targeting was based on indicators derived from 2001 Census data. For 2011 a more sophisticated algorithm would be developed. Additionally, in order to get comparability with 2001, random swapping plus SCA was looked at. There was therefore comparison between a total of six different methods of disclosure control. Three different levels of perturbation, 2, 10 and 20 per cent, were used at first, but following advice from the working group, the final analyses concentrated on 2 per cent perturbation.

(i) **Quantitative evaluation** was carried out by applying these six methods to two sets of tables from the 2001 Census, one set from enumeration area (EA) SJ, which is a mainly urban area including Southampton, the other set from EA KB, which is a more rural area of Cheshire. The disclosure risk and utility of the resulting tables were calculated. The test data together with the risk and utility measures are described in detail in Appendix B.

Disclosure risk was measured using specially written SAS programs, which looked at small cells, disclosure by differencing and different types of attribute disclosure. Utility was measured by the in-house package Infloss, which is also based on SAS. This looked at the effects of the perturbation on various properties of the tables, such as totals and subtotals, distortion to distributions, and impact on variance and measures of association. The results of the quantitative evaluation are given in Appendix C.

One type of table looked at was origin-destination, and it was found that none of the disclosure methods were really satisfactory without having a serious detrimental effect on the utility of the more detailed tables. The working group accepted the recommendation that origin-destination tables should generally

be protected by licensing and restricted access.

(ii) **Qualitative evaluation** was based on a set of twenty criteria, four mandatory and sixteen secondary, which were agreed by the working group. The criteria were weighted according to how important it was considered to be that the chosen disclosure method should satisfy them. Each method was scored against each criterion, and the totals were calculated. The criteria are described fully in section 4, and the scores are given in Appendix D. The scores were based on the results of the quantitative evaluation, together with further information obtained during the investigations and discussion with the working group.

There were originally fourteen secondary criteria. The UKCDMAC suggested two more, when they peer-reviewed the evaluation. These additional criteria also related to aspects which were raised by members of the ONS Statistical Policy Committee (SPC) and UKCC. These are highlighted in Appendix D.

(4) Overview of disclosure methods

The three short-listed methods are described in detail in Section 2 and Appendix A.

(i) **Record swapping** is a pre-tabular method of perturbation. A small percentage of households are selected, either randomly or by targeting households which are considered to pose particular disclosure risk. For each of these households another household with similar attributes is found, generally within the same local authority district (LAD). The geographies of the two households are then swapped. This technique is described fully in sections 2.1 and A.1 in Appendix A.

(ii) **Over-imputation** is also a pre-tabular method, which was implemented by using CANCEIS (Canadian Census Edit and Imputation System). This tool is used by ONS as the edit and imputation tool for correcting missing or inconsistent raw census data. For over-imputation, a small percentage of households are first selected, either randomly or targeted. For each of these, variables corresponding to particular attributes of each member of the household are blanked out. For each of these variables, CANCEIS then finds the best possible match for the household, generally within the same LAD. The variables for each person in the matched household are then copied into the records for the selected household. The method of over-imputation is described fully in sections 2.2 and A.2.

In the first stage of the evaluation, geography and age were the variables chosen for over-imputation for the SJ tables, and geography alone for the KB tables. The Working Group recommended that the over-imputation work should be repeated, but using only non-geographical variables. There was therefore a second stage to the evaluation, in which *non-geographic* over-imputation was examined.

(iii) **IACP** is a post-tabular method. All the cells in a table are perturbed by adding a small positive or negative number, which may be zero. These perturbations are based on keys assigned to each record in the underlying microdata. There is an outline description of the technique in section 2.3, and a full description of the method and its application is given in A.3.

(iv) **Small cell adjustment**, which was included to enable comparability with 2001 Census, is a post-tabular method. Small cells in each table are randomly perturbed, using an unbiased probability scheme. This is described in section 2.4.

(5) Peer review

Members of UKCDMAC made a number of comments at various stages of the evaluation and raised some concerns. These are addressed in section 3.1.1..

(6) Reason for recommendation to use record swapping

The working group agreed the wording and weighting of the criteria. The group also agreed the scores for the four mandatory (M1-M4) and 14 secondary criteria (S1-S14). The two additional criteria, S15 and S16, with their weights and scores, were agreed by the UKCC and SPC. See section 3.2 and Appendix D for details. The total scores were:

	Record swapping	Over-imputation	IACP
Mandatory criteria	200	200	180
Secondary criteria (original)	340	358	248
Secondary criteria (including 2 more suggested by UKCDMAC)	403	386	311
Total	603	586	491

The scores for record swapping and over-imputation were not significantly different while the IACP method scored significantly lower since the method does not maintain complete consistency between tables. Both record swapping and over-imputation would be able to manage the risk of disclosure and disclosure by differencing. Hence the choice between them is made on the impact of each method on the data utility.

Detailed discussions at the working level concluded with agreement that the weaknesses of record swapping could be overcome through careful design, whereas the weaknesses of over-imputation were considered implicit to the method and more difficult to overcome.

The weaknesses of over-imputation are that, at the levels of perturbation assessed, (i) the method distorted associations between variables (Criterion S3), (ii) impacted on totals and sub-totals within tables at all geographies

(though it does not affect the total number of individuals in any geographical area) (Criterion S6) and (iii) it has not been implemented satisfactorily in tests (Criterion S14). Currently there is no accepted methodology for over-imputation for SDC, as CANCEIS is designed to insert values as near as possible to the true values of variables, rather than to change true values to different ones. Additionally the fact that legitimate data items are removed and replaced with imputed values was considered to be unpopular with users (Criterion S15). There will also be some outputs, including those at small geographies, where over-imputation could not be applied, since not every variable could be satisfactorily imputed on. For example if over-imputation were applied to sex, marital status, ethnic group or religion then either this would create difficulties in maintaining consistency with other variables or else it would be very likely that the real value would be imputed (Criterion S16).

The weaknesses of record swapping are that (a) it could be possible to match high level tables against microdata samples and determine and locate population uniques (Criterion S12) and (b) it would be more difficult to protect special populations such as communal establishments and workplaces (Criterion S8). However, it would be possible to address these issues (a) predominantly through licensing arrangements and (b) through careful design of the record swapping methods. It is also more difficult for record swapping to take into account the data quality of different variables (Criterion S4) but it could consider the data quality related to response rates and response-related imputation.

The key strength of record swapping over over-imputation is that no persons or data items are removed from the census data and therefore outputs at national level and high geographies will be unaffected by record swapping. Record swapping has also been used before (in the UK and USA) to protect census tables, whereas over-imputation has not. However the method of record swapping for the 2011 Census will differ in several ways from that used for the 2001 Census. For example it will target "risky records" rather than selecting records at random. Since the targeting algorithm will be more sophisticated than that used in this evaluation, and since there will be other considerations such as population thresholds, the record swapping methodology will result in a much lower disclosure level than is indicated by the results in Appendix C. Thus the data will be sufficiently protected by the method of record swapping which will be applied, without there being any danger of needing additional post-tabular protection, as happened last time.

(7) Conclusion

The recommendation to use record swapping as the primary disclosure control method for 2011 Census, supported by this evaluation, was presented to the ONS Statistical Policy Committee in September 2009, and was approved. The UK Census Committee, which represents the RsG, also approved the recommendation.

Further work will be necessary to establish the details of how record swapping

would be implemented, including levels of swapping, approaches for targeting and taking into account imputation due to non-response. In particular a new algorithm for targeting will need to be developed. Record swapping will be used in conjunction with population thresholds and the level of detail made available in outputs, taking into account any special treatment which might be considered necessary for more sensitive variables.

Summary

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. In July 2007 a review of a wide range of SDC methods was undertaken assessing them against a set of qualitative criteria in line with the policy statement made by the Registrars General. This resulted in three SDC methods being short-listed for further evaluation to assess risk and utility quantitatively, and this short-list was agreed by the UK Census Committee (UKCC) including the Registrars General. This paper provides an evaluation of the three methods. The approach to this evaluation has been reviewed by the UKCDMAC SDC subgroup and agreed with the UK SDC subgroup.

The evidence provided in this paper has been used to inform a recommendation for the SDC method to be used for 2011 Census tables. This has resulted in agreement that record swapping will be the primary strategy for disclosure control.

1. Background

Census output can be released in a number of different formats; standard pre-planned tables, commissioned tables requested by users, user defined tables via flexible table generating software and census sample microdata. Publishing aggregate or individual data carries the risk that individuals or entities could be identified and confidential information about them could be released. The UK Census Offices need to protect the confidentiality of census respondents for a number of reasons. The production and use of official statistics depends on the cooperation and trust of citizens. Such trust cannot be maintained unless the privacy of individuals' information is protected. There are also legal and policy obligations that must be respected.

The aim of statistical disclosure control (SDC) is to ensure that statistical outputs provide as much value as possible to users while protecting the confidentiality of information concerning individuals or entities. SDC methods modify or summarise the data and there is a range of different methods that can be used to protect census outputs. SDC methods can be pre-tabular (applied to the underlying census records) or post-tabular (applied to tables). This paper focuses on the work that has been undertaken to develop an SDC strategy for tabular outputs for the 2011 Census.

For the 2001 UK Census the initial plan was that tables would be protected by a pre-tabular method of disclosure control, namely random record swapping. This method of disclosure control was followed up by applying population thresholds to the tables. Following a review, the Office for National Statistics (ONS) and the Northern Ireland Statistics and Research Agency (NISRA) decided to adopt larger thresholds than those previously agreed with the General Register Office for Scotland (GROS). Prior to releasing tabular outputs from the 2001 Census, concerns were raised that the public would perceive that no disclosure control method had been applied. ONS decided

that the additional method of small cell adjustment was required for tabular outputs. The small cell adjustments added more uncertainty and removed small cells from tabular outputs. NISRA also applied the additional method of small cell adjustment but GROS did not. This late change in SDC methodology and lack of UK harmonisation caused a number of problems for users.

In November 2006 the UK SDC Policy position (ONS (2006)) for the 2011 Census was agreed by the Registrars General of Scotland, England and Wales and Northern Ireland. The Registrars General agreed to aim for a common UK SDC methodology for 2011 Census outputs to achieve harmonisation. The SDC Policy position is based on the principle of protecting confidentiality set out in the National Statistics Code of Practice (which has now been replaced by the UK Statistics Authority Code of Practice for Official Statistics, see below). The Registrars General concluded that the Code of Practice guarantee of confidentiality can be met in relation to census outputs if no statistics are produced that allow the identification of an individual (or information about an individual) with a high degree of confidence. The Registrars General considered that, as long as there has been systematic perturbation of the data, the guarantee in the Code of Practice would be met. It was therefore agreed that small counts (0's, 1's, and 2's) could be included in publicly disseminated census tables provided that

- a) uncertainty as to whether the small cell is a true value has been systematically created, and
- b) creating that uncertainty does not significantly damage the data.

The decision to allow small cells in publicly disseminated tables means that both pre-tabular methods and post-tabular methods or combinations of the two can be considered for 2011. The Registrars General have expressed a preference for pre-tabular methods, provided there is not undue damage to the data, and have also stated that the key risk is attribute disclosure. The exact threshold of uncertainty required has not yet been decided but the RsG have stated their preference for SDC to have a 'light touch'.

Since agreeing the UK SDC Policy, the Statistics and Registration Service Act 2007 (SRSA) has come into force. Section 39 (2) of the Act defines personal information as information which relates to and identifies a particular person. It specifies what constitutes a disclosure of information and the sanctions that may apply for any breach of confidentiality.

Disclosure of personal information can take place through being specified in the information, by being deduced from the information, or by being deduced from the information when taken together with any other published information (Section 39 (3)). The 2007 Act states that personal information must not be disclosed unless through an exemption as specified in Section 39 (4). The UK SDC Policy is in line with Section 39 of the SRSA. Following the introduction of the SRSA the National Statistics Code of Practice has been superseded by the Code of Practice for Official Statistics (January 2009), which affirms as one of its eight fundamental principles that 'Private information about

individual persons (including bodies corporate) compiled in the production of official statistics is confidential, and should be used for statistical purposes only'. The new code provides a similar level of confidentiality protection as the old code and the UK SDC policy conforms to the new code.

In July 2007 a review of a wide range of SDC methods was undertaken, assessing them against a set of qualitative criteria in line with the policy statement made by the Registrars General, ONS(2007) . This resulted in the following three SDC methods being short-listed for further evaluation to assess risk and utility quantitatively:

- Record swapping
- Over-imputation
- ABS Cell Perturbation Method (developed by the Australian Bureau of Statistics)

The previous report ONS(2008) circulated in October 2008 provided an evaluation of these three methods benchmarked against the method used in 2001 (small cell adjustment with record swapping). Comments from both the UKCDMAC SDC sub-group and the ONS SDC Working Group meant that further work was needed. This paper includes the additional work and, together with the content of the previous report (most of which is incorporated here), was used to inform a recommendation for the SDC method to be used for 2011 Census tables.

The next section of this paper provides an overview of the methods that have been evaluated. Section 3 describes the approach used to evaluate the methods both quantitatively and qualitatively and the criteria which were used to assess the results. Sections 4 summarises the results and draws conclusions from them, including the final sign off by the RsG. Section 5 looks at the next steps. More detailed analysis and technical details are given in Appendices A, B and C. Appendix D shows how each disclosure control method measured up against the assessment criteria, and Appendix E contains a glossary of terms.

2. SDC methods

This section provides a high level description of the three short-listed methods and the method used in 2001. More details are provided in Appendix A.

2.1 Record swapping

Record swapping involves perturbing the data by swapping the geographical identifiers of a small percentage of household records with other records, matching on specific control variables (e.g. age, gender, Hard to Count (HtC) index¹, household size). Swapping only records which match on control

¹ The Hard to Count (HtC) index was constructed in the 2001 UK Census as a measure of enumeration difficulty. It was constructed from the following 1991 Census variables; Multi-occupancy, unemployment, language difficulty, private rented accommodation, number of household imputed in

variables helps to minimise bias. A small percentage of individual records within communal establishments can also be swapped using similar control variables but replacing household size by communal establishment type. Record swapping would generally be carried out within a local authority district (LAD) and households / persons in communal establishments are swapped in and out of smaller geographical areas e.g. output areas (OAs)². If a match can be found for every record selected, record swapping ensures that local authority marginal distributions remain unaffected. However, in order to protect very unusual households where a match cannot be found, some records are swapped across LAD boundaries, so some LAD counts will be slightly affected. This across-LAD boundary swapping would meet the requirements expressed for the 2001 Census by Dick Carter of Statistics Canada who was commissioned to review arrangements for that census.

Record swapping can take different forms including random record swapping or targeted record swapping. Random record swapping involves selecting, at random, households and individuals within communal establishments for swapping. Targeted record swapping involves selecting a random sample of the potentially unique/ risky records for swapping. This generally involves flagging records which are considered to be risky based on particular characteristics. The records actually chosen for swapping can then be selected from these flagged records. The selected records are then paired for swapping with other flagged records which match on control variables. For records where no match can be found within the flagged records, a match is found using non-flagged records which match on control variables. Appendix A, section A.1, describes in more detail how record swapping was used in the evaluation exercises.

2.2 Imputation

Imputation is a commonly used method for replacing missing values in census and survey data due to item non-response. A new method of over-imputation has been devised for disclosure control of census data using CANCEIS (a specially designed package developed by Statistics Canada).

In the earlier stages of the evaluation, work concentrated on geographic imputation, to attempt consistency with record swapping. Geography and age were blanked out and imputation used as a method of disclosure control, i.e. over-imputation. Donors from the remaining population were used to replace values. In the case of imputing geography; enumeration districts (EDs) or OAs were imputed within the same LAD. Age was imputed using all possible donors.

However, the method of imputation that is evaluated in this report is to impute variables other than geographic ones. One repeated comment from both the

1991. Scotland also used ethnic group.

² In 2001, Output Areas were the smallest geographic building block, that could be combined to form higher geographies such as Local Authorities.

UKCDMAC SDC sub-group and the UK SDC Working Group was to look more closely at non-geographic over-imputation since this was likely to provide better protection for both tables and microdata. As with geographic imputation, donors are found from the remaining population that match or closely match on other related variables. CANCEIS will attempt to find the best possible match, firstly matching on 35 specified variables if possible, then failing that on a subset of them. CANCEIS might not necessarily match on all variables for a given recipient/donor pair, but uses a probabilistic method of choosing which combination of matching variables is best in a given case. There is also a distance function specified to find or prioritise donors within certain geographical limits of the recipient.

As with record swapping over-imputation can either be applied to a random sample of records or specific high risk records can be targeted for imputation. Section A.2 describes in more detail how imputation was used in the evaluation exercises.

2.3 ABS cell perturbation

This new cell perturbation method developed by the Australian Bureau of Statistics (ABS) is essentially a post tabular approach which takes into account pre-tabular information. The method involves adding small perturbations³ to all cells in a table using a two stage process. Stage one results in a consistently perturbed non additive table. At the second stage another perturbation is added to each cell (excluding the grand total) to restore table additivity.

In stage one all microdata records are assigned a record key. When creating a table the record keys for all records contributing to each internal cell are summed and a function is applied to this sum to produce the cell key. Lookup tables (determined by the organisation) are then used where the true cell value and the cell key are used to determine the amount by which the cell count should be perturbed. This means that the same cell is always perturbed in the same way. The perturbation can be set to zero for a pre-determined set of key outputs (e.g. age by sex population counts). Table margins are perturbed independently using the same method.

The stage two perturbations are generated using an iterative fitting algorithm which attempts to balance and minimise absolute distances to the stage one table, although not necessarily producing an 'optimal' solution.

For this evaluation a modification of the original ABS method, referred to as Invariant ABS Cell Perturbation (IACP), has been developed to improve utility by attempting to make the first stage perturbations invariant with respect to the table cell frequencies⁴. Full details of IACP are given in section A.3.

³ Note that some perturbations will be zero

⁴ There may be a small increase in the number of zeros in tables when applying IACP

2.4 Small cell adjustment (SCA)

Applying small cell adjustments involves randomly adjusting small cells upwards or downwards to a base using an unbiased prescribed probability scheme. Marginal totals are obtained by summing perturbed and non-perturbed cells. Small cell adjustments were used in addition to random record swapping to protect 2001 Census tabular outputs for England and Wales and Northern Ireland. In Scotland SCA was only applied to tables counting persons by workplace, because record swapping did not protect this output.

3. Approach

This section describes the approach used to evaluate the short-listed SDC methods. The main aim of the disclosure control strategy is to reduce disclosure risk to an acceptable level whilst maintaining as much data utility as possible. Quantitative risk and utility measures were evaluated for different 2001 Census tables, as described in 3.1. SDC methods have qualities which cannot be accounted for quantitatively and thus the qualitative advantages and disadvantages of the methods must also be addressed. These criteria are described in Section 3.2.

3.1 Evaluating risk and utility quantitatively

The most important characteristic of the SDC strategy for 2011 Census is that disclosure risk should be managed to an acceptable level, in order to respect legal and policy obligations and to ensure public co-operation and trust are maintained. Census outputs have a higher risk of disclosure and are harder to protect than other statistical data outputs because they contain whole population counts, because small areas predominate in output geography, and because tables are disseminated from only one data source, so that tables can be linked and differenced.

The Registrars General have highlighted that the key disclosure risk for 2011 Census output is attribute disclosure, i.e. learning something from the census data about an individual or group of individuals that was not previously known. Attribute disclosure is highly associated with – but not exclusively so – low numbers in tables. If an intruder knows something about a person e.g. which row the person will be counted in, then he will deduce something new about the person if all the cases in that row are in one column and all other columns in the row contain zero. Such situations are certain to arise when there is only one case in the row, and more likely to occur when the number of cases in the row is small. Hence the analyses in this paper concentrate on the effects of possible SDC methods on the numbers of cells with values zero, ‘1’ and ‘2’ in tables.

As described in the previous paragraph, attribute disclosure occurs if there is a row or column that contains mostly zeros and a small number of cells that are non-zero. One can then learn a new attribute about an individual or a group of individuals. Three different measures for attribute disclosure are considered. Group disclosure occurs when all respondents fall in one cell in a row (or column). Negative attribute disclosure occurs when rows (or columns) contain only zeros. Within-group disclosure occurs when there is a single respondent in a cell in a row (or column) where all other respondents fall in another cell. One could say this is a special case of differencing (see below); the two tables concerned are the one published and a notional table in the mind of the intruder containing one case – which he subtracts from the published table. The different kinds of attribute disclosure are described in more detail in Appendix B, section B.2.

The evaluation also considered disclosure risk due to small cells, see B.2.4 and the risk of disclosure by differencing, see section B.2.5. It is vital that the SDC method selected provides protection against disclosure by differencing and linking (both for geographical and other variables). Protecting against disclosure by differencing and linking increases the flexibility of outputs in general and removes the need for auditing ad-hoc outputs, e.g. commissioned tables, which can be resource intensive.

Managing risk will necessarily impact on data utility and the aim is to adopt an SDC strategy that minimises this effect, while still providing effective risk management. For this evaluation, data utility is measured against the following requirements:

- i) All tables should be additive (i.e. rows and columns add up to row and column totals)
- ii) All cells should be consistent across tables (i.e. the same cell in a different table has the same value); note that inconsistencies (the number and nature of them) may in themselves be disclosive both about the subjects of inconsistent cells and about specific details relating to the method of disclosure control.
- iii) Relationships between variables should be maintained as much as possible
- iv) The method should be unbiased
- v) The method should have a minimal impact on cell values, particularly totals and subtotals
- vi) The method should have a minimal impact on the variance of the estimates
- vii) The method should have a minimal impact on statistical analyses

The trade off between risk and utility is evaluated quantitatively. Unperturbed 2001 Census microdata were obtained for two Estimation Areas (EAs): SJ (Southampton, Eastleigh and Test Valley districts) and KB (Congleton, Chester, Crewe and Nantwich, Ellesmore Port and Vale Royal districts). KB is a rural area chosen for the sparsity of its population whereas SJ is more urban and densely populated. For the pre-tabular methods 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. For the IACP method the microdata keys were assigned to the individual records and then perturbation applied once the table had been created, see section A.3. Risk and utility measures were calculated by comparing the original and protected tables for each SDC method.

3.1.1 The two stages of the evaluation

The quantitative analysis has taken place in two stages. The first compared geographic over-imputation, record swapping and IACP across three levels of

perturbations (2, 10 and 20 per cent) with the two pre-tabular methods also broken down by whether perturbed records were selected at random or whether targeted. The three short-listed methods were thus broadly comparable. Note, 2 per cent perturbation for swapping/over imputation results in 2 per cent of *records* being perturbed, whereas 2 per cent perturbation for the IACP method results in 2 per cent of *cells* in the table being perturbed. For this evaluation it was assumed that no imputation for non-response had been applied, to keep the analysis more straightforward. However, the potential protection from imputation for non-response should be considered when refining the chosen method for 2011. Note, it would be possible to take this into account for pre-tabular but not for post-tabular methods. We have also not taken into account any protection through the natural uncertainty that might be introduced by informing users that disclosure control methods had been employed (precise details of the chosen method, such as the levels of perturbation etc., would not be released).

Risk and utility measures were evaluated for the following tables, each with different characteristics:

For EA SJ:

Table 1: Country of birth by sex by religion

Table 2: Number of persons in household by accommodation type

Table 3: Age by sex by marital status

Table 4: Origin-destination table

For EA KB:

Table 5a: Age by ethnic group by sex for all persons

Table 5b: Age by ethnic group by sex for persons without limiting long term illness

Table 5c: Age by ethnic group by sex for persons with limiting long term illness

In the case of EA SJ, the data used for assessing over-imputation were slightly different to the files used for the other methods. The geography definitions on the file were not identical and thus slightly different tables were analysed (this was a side-effect of using CANCEIS which would need to be addressed if over-imputation were selected as the preferred method).

A report produced on the basis of the above analysis was presented to both the UKCDMAC SDC sub-group and the UK SDC Working Group (which includes members from the four UK Census Offices), ONS (2008). Though the results were helpful, both groups commented on the need for further analysis. The comments meant that a second stage of the evaluation was deemed necessary, the key aspects being to:-

- restructure the tables to ensure a far greater number of instances of attribute, group and within-group disclosure and provide a more well-

founded analysis of disclosures that are likely to arise during processing 'live' Census tables. In effect, since each row corresponded to one geographical area, the analysis in the first stage had highlighted only those instances where there were zeros across all, or all but one cells, across any one geographical area, or where non-zero counts only occurred in the corresponding cells of one or two areas, where cells represent all possible cross classifications of the variables spanning the table. We have now considered the variables, rather than geography, as the rows and columns of tables taking the cross classification of variables two at a time, e.g. for Table 1 we consider instances of attribute disclosure for sub tables of country of birth x sex, country of birth x religion, and sex x religion, for each ward in the EA.

- assess *non-geographic over-imputation* as a method for protecting against disclosure. In the first stage, we had imputed geography (either ward or OA) in order that every table at that level would have some protection. An alternative view was to impute on the variables within the table where initial analysis had proved promising.
- concentrate in this second stage on small perturbation levels – 10% and 20% were useful in highlighting broad strengths and weaknesses of the competing methods but were felt to be unrealistic in terms of the level that might be employed in practice. Fresh analysis would therefore concentrate on the 2% perturbation level.

Subsequent to the first stage analysis being completed, some errors were found in the microdata used for swapping and imputation, so that, given the time and available resource, we have concentrated within the second stage on Tables 1 and 3 for SJ EA only. Tables 5A-C, using KB EA have been used from the first stage to assess consistency, additivity and disclosure by differencing.

More details on the data used and risk and utility measures are provided in Appendix B.

3.2 Evaluation

At the UK SDC Working Group meeting in Belfast 17 November 2008 it was decided that some further analysis on the short listed methods was required, after which a final recommendation would be made. This resulted in agreement on a set of criteria against which the results would be assessed. The following table lists the criteria to be used to score the methods and gives a weighting to each criterion to reflect its importance. The list builds upon the initial assessment criteria used to construct the shortlist of methods, with the extra criteria reflecting some of the issues raised since the analysis was started and to allow greater visibility of the strengths and weaknesses of the different methods. Four mandatory and 14 secondary criteria were agreed by the working group. The UKCDMAC SDC sub-group, when peer-reviewing the evaluation, suggested two more.

Note that the assumption is made that the dataset being used is complete, with no errors introduced by either respondent capture, edit or imputation. This assumption is unrealistic for live running, but it is made for the purpose of comparing methods in this evaluation.

The criteria are scored on a scale of 0-5.

- 0: The criterion is not met at all
- 1: The criterion is partly met, but only to a very limited degree
- 2: The criterion is sometimes met, or to some degree
- 3: The criterion is usually met
- 4: The criterion is nearly always met, or almost completely met
- 5: The criterion is always met

In addition the criteria were given weighting factors, with the mandatory criteria each having a weighting of ten. There is also a requirement for the method to satisfy or mostly satisfy all mandatory criteria (i.e. a score of at least 4) in order to be considered for the final choice of method. The total score for each method is calculated as Σ (Weighting x Score) over the criteria.

The following are the criteria used. Section 4 describes these in more detail and relates them to the results from the quantitative evaluation.

MANDATORY CRITERIA

<u>Label</u>	<u>Description</u>	<u>Weighting</u>
M1	The method creates the desired level of doubt about any attribute disclosure and protects against differencing	10
M2	Marginal totals in protected tables are unbiased	10
M3	Protected tables are additive	10
M4	The method cannot be unpicked	10

SECONDARY CRITERIA

<u>Label</u>	<u>Description</u>	<u>Weighting</u>
S1	Method provides consistent cell counts and totals between different protected tables	9
S2	The method is practical bearing in mind the resources available in terms of manpower, computing power and software costs	8
S3	For a given level of risk relationships between variables are maintained in protected tables	7
S4	The method can take into account the levels of imputation and overall data quality of different variables	6

S5	Counts of households and residents for small areas are not unduly perturbed	6
S6	The method does not unduly perturb/affect counts for large geographies (e.g. LA level and above)	6
S7	The method has a low impact on the variance of estimates	6
S8	The method can be used or adapted to protect outputs from special populations such as communal establishments or from workplaces	6
S9	Will not restrict the detail of releases or the subsequent protection method to be used for microdata samples	6
S10	The method and any required software will have adequate lifespan for purpose	6
S11	The method can easily be accounted for by users in analysis	5
S12	The same method can be applied to microdata outputs	5
S13	The method is likely to be easily understood by users	5
S14	The method has been effectively used for protecting similar outputs	4

UKCDMAC suggested two additional secondary criteria:

S15	The method makes use of all data collected in the Census	7
S16	The method will be applied systematically to all tables and all cells	7

4 Summary of results

This section provides a high level overview of the performance of each SDC method against the criteria listed in Section 3. Detailed results from the quantitative analysis are provided in Appendix C but this section provides a summary of the key findings that relate to each of the criteria.

4.1 Mandatory evaluation criteria

4.1.1 M1 The method creates the desired level of doubt, measured in terms of protective changes to tables, about any attribute disclosure and protects against differencing

The level of protection provided by all three short-listed methods is restricted by the low level of perturbation used for the evaluation (2 per cent). Imputation and swapping each offer some protection, with swapping better than imputation in many cases. IACP leaves all zeros unchanged, though there would be some uncertainty as to whether zeros in the protected table were real zeros, since they could have been perturbed from non-zero cells. Much of the protection in practice for all three methods would be through user perception that there has been some disclosure control employed, without users being made aware of the full details.

Over-imputation could not be applied to all variables. There is therefore the possibility that a table based on variables which were not selected for imputation may not have any disclosure protection (other than edit and imputation and the perception of disclosure). However, this risk should be relatively low since most key variables would be imputed, see 4.4.3.

At higher geographical levels (above LAD) record swapping will provide little protection from attribute disclosure since swapping is mostly, but not entirely, carried out within LADs (see Section 2.1). However, as the geographical size of an area increases, the disclosure risk decreases.

Each of the three methods provides some protection against disclosure by differencing. There is always ambiguity over any cell value and hence there will always be ambiguity over the true value of a cell that is derived from differencing two tables. Further work would be required to examine differencing issues where tables are produced for two non-coterminous geographies, e.g. ward and output area, to assess the risk from small slivers between the two geographies. This would need to be considered for any of the short-listed methods. For record swapping with SCA, if the level of swapping is low, there would be problems with differencing between combinations of larger cell values (that will have not been perturbed by the SCA approach).

4.1.2 M2 Marginal totals in protected tables are unbiased

The IACP method can be made unbiased by carefully designing the look-up

table. Marginal totals from the imputation and swapping methods should be unbiased.

4.1.3 M3 Protected tables are additive

For pre-tabular disclosure control methods (record swapping and over-imputation) all tables will be generated from the perturbed microdata and hence will be fully consistent and additive. For SCA and the IACP method the tables will be additive but not fully consistent.

4.1.4 M4 The method cannot be unpicked

Despite the likely difficulty in users understanding the IACP method, it remains possible that the method could be unpicked. Although the look-up table may remain confidential, distributions of different values of cells with supposedly the same value may be helpful to an intruder. As long as details of the swapping and imputation methods are not disclosed, it is unlikely that users could establish the precise method, variables perturbed (in the case of over-imputation) or the level of perturbation. The lack of protection for zeros with the IACP method could mean that it is easier to unpick than the other two methods.

4.2 **Secondary evaluation criteria**

4.2.1 S1 Method provides consistent cell counts and totals between different protected tables

Since record swapping and over imputation are pre-tabular there is no difference between a table that is directly extracted or one that is produced by differencing two protected tables, therefore both these methods preserve consistency.

For post-tabular methods, producing a table by differencing two protected tables will sometimes produce different results than extracting the table directly and then applying the SDC method.

The results for Table 5C show that the IACP method is not consistent. Where a table is produced by differencing two protected tables this is different to the table that is extracted and protected directly, e.g. for 90 per cent IACP, 3.8 per cent of cells had different values. The differenced table can have negative values and we have observed small differences in a small number of row/column totals.

4.2.2 S2 The method is practical bearing in mind the resources available in terms of manpower, computing power and software costs

The SDC methods selected for the 2011 Census should be practical, easy and quick to implement. This will minimise the risk of errors and facilitate the release of outputs in a timely manner to an agreed timetable. Any SDC

method employed should not affect timeliness by requiring excess computer run-time each time the method is used to protect a table (this could be a particular issue for post-tabular methods).

Pre-tabular methods are in general easier to implement than post-tabular methods (particularly if tables are generated on-line) since they only need to be applied to the microdata once and then all tables are to be generated from the perturbed microdata.

Record swapping was used in 2001 and is fairly straightforward to implement. Although yet unproven the implementation of over-imputation could utilise the CANCEIS software already in use at the ONS for edit and imputation and would therefore be relatively straightforward. However, edit checks would be required when implementing over-imputation to ensure that no illogical records were created in the process.

The ABS perturbation method has been implemented by the ABS for their 2006 Census, but at present this is only used for standard tabular outputs. Although it is reported that the method works efficiently, the functionality for use with flexible table generating software is not yet proven. The ABS has merged the SDC method with the table building software SuperCross for their implementation. Within this evaluation the perturbation method has been programmed in SAS and runs quickly, but it has not been tested thoroughly in a real-time environment. In particular the IACP method involves procedures to calculate the transition matrix, which has to be run each time a different table is generated.

Record swapping was used in conjunction with small cell adjustment in England and Wales and Northern Ireland in 2001. Record swapping was used without small cell adjustment in Scotland.

All methods are flexible in that the amount of perturbation applied to the data can be determined. The pre-tabular methods can be either random or targeted. Before implementation, decisions would be needed on which variables should be imputed in non-geographical over-imputation. In this evaluation all variables spanning the tables have been selected for potential imputation. This would not be feasible when considering all 2011 Census tables. Key variables for imputation would need to be determined and this would mean that tables not involving these key variables would have no protection other than perception. The IACP perturbation method is the most flexible in that the look-up table can be determined for each output table to control the level and distribution of perturbation for different cell values – at the expense of increased inconsistency.

4.2.3 S3 For a given level of risk relationships between variables are maintained in protected tables

At the individual level record swapping has no impact on the relationships between variables since only geography is swapped between households. The results for the Cramer's V test show that over-imputation does have an

impact on the relationships between variables. This could potentially impact on the relationship between individuals in households where some variables for some individuals will be imputed and others will not.

4.2.4 S4 The method can take into account the levels of imputation and overall data quality of different variables

An advantage of pre-tabular methods is that one can take into account imputation for non-response when implementing the method, i.e. one could reduce the level of perturbation (swapping or over-imputation) in areas where ambiguity had already been introduced by non-response imputation. For non-geographic imputation, there is a possibility of taking into account non-response for specific variables. For post-tabular methods it is not possible to take into account levels of imputation for non-response ..

4.2.5 S5 Counts of households and residents for small areas are not unduly perturbed

For record swapping counts of households and residents will not be affected for small areas. Over-imputation will also not affect any counts since only the characteristics of those households and residents are perturbed. There will be small differences in the numbers of households and residents at small areas for the IACP method since it is a post-tabular method.

4.2.6 S6 The method does not unduly perturb/affect counts for large geographies (e.g. LA level and above)

Record swapping has no effect on counts for large geographies, unless no match can be found within the LA geography and a match is needed from outside the district. IACP has some effect on counts at all geographies but over-imputation does affect counts more, particularly where selected variables are difficult to impute accurately as shown in the results for the distance metrics.

The results show that record swapping does not impact on totals and subtotals across the tables but the IACP and over-imputation methods do, e.g. the protected table could have more men and less women than the original table.

4.2.7 S7 The method has a low impact on the variance of estimates

Swapping and IACP generally maintain the variances while imputation has a slightly larger effect. This is particularly so in the targeted imputation where 'risky' records that might be on the edge of distributions may have values imputed from donors that are closer to the centre of distributions. All methods will have an effect on the variance of geographical indicators such as the percentage of persons of pensionable age in an OA. For example, record swapping has the effect of homogenising the populations across OAs and hence of reducing the variance of geographical indicators.

4.2.8 S8 The method can be used or adapted to protect outputs from special populations such as communal establishments or from workplaces

IACP could be used to protect these outputs whereby all cells are susceptible to perturbation, so modifying the counts and characteristics of persons in workplaces and communal establishments. Neither record swapping nor imputation could protect the counts of persons in workplaces or communal establishments, but imputation could be used to modify the characteristics of the persons in both, while swapping could swap the characteristics with those of other persons. Over-imputation would also be easier to implement for these tables in comparison to swapping. For some individuals it may be difficult to find suitable matches in other communal establishments/workplaces. Swapping individuals between workplaces could potentially distort tables that combine residence and workplace. All methods would need to be used in conjunction with thresholds in order to fully protect individual establishments or businesses (although they are not data providers); workplace zones need to be designed with this in mind.

4.2.9 S9 Will not restrict the detail of releases or the subsequent protection method to be used for microdata samples

Although the focus here is on SDC methods for tables we also consider the impact (if any) on microdata outputs.

The impact of the SDC methods for tables on microdata and the interaction between different types of output in terms of linking should be given consideration when short-listing. Methods which leave a high proportion of true '1's and '2's in tables could impact on microdata releases, since one could use the tables to determine or locate population uniques in microdata samples.

All methods other than SCA will leave some true small cells in protected tables so there will be a risk that microdata could be combined with released tables to determine sample uniques in the microdata. This risk is greater for record swapping in comparison to non-geographic over-imputation since population uniques at the national level will not be perturbed, i.e. a '1' in a national table will be a true '1'. An intruder finding a unique record in a microdata sample could, by matching to a national table, deduce that the record is a true population unique. In 2001, since GROS did not apply SCA, this issue resulted in limitations on the release of microdata samples in Scotland and additional resource intensive checks being made. However, the format of the 2011 Census microdata samples are yet to be determined and, if applying record swapping protection, will be provided against this risk by perception as well as by any licensing restrictions. There would be greater protection against '1's at national level if over-imputation were employed, since there would be some possibility that one or more of the characteristics pertaining to the '1' had, in fact, been imputed.

4.2.10 S10 The method and any required software will have adequate lifespan for purpose

The SDC methods recommended for 2011 should be future-proof so that implementation throughout the life cycle of the census data is possible (future-proofing should also be considered with respect to dependency on any software required to implement the methods). Record swapping can easily be programmed in any statistical software language, e.g. SAS. Over-imputation relies on support for CANCEIS but once done will be future-proof. The IACP method could potentially rely indefinitely on the chosen tabulation tool.

4.2.11 S11 The method can easily be accounted for by users in analysis

It will be important to provide users with information on how the SDC methods may impact on their analyses and how this impact can be taken into account. Since the level of perturbation is not revealed in any of the short-listed SDC methods it would not be easy for users to account for the impact of the method in their analysis although some information could be supplied in the metadata. Record swapping will not impact greatly on analysis for LA levels and above.

4.2.12 S12 The same method can be applied to microdata outputs

Record swapping does not provide any protection to microdata since geography is swapped at geographical levels that are lower than that likely to be released within microdata samples. Over-imputation will provide some protection when variables other than geography are imputed, however it is likely that further perturbations may be required to fully protect a microdata release. The IACP perturbation method and SCA are post-tabular and hence provide no protection to microdata samples.

4.2.13 S13 The method is likely to be easily understood by users

To achieve user acceptance it will be important to keep the SDC method simple and easy to understand, while ensuring that it is not possible to unpick it.

Record swapping is a widely accepted method of data protection and is simple to understand. User acceptance is thought to be lower for over-imputation than record swapping since it is less well known, original data values are sometimes deleted and the detail of the method is more difficult to understand. The IACP perturbation method is complex and is not easily described or understood by users.

4.2.14 S14 The method has been effectively used for protecting similar outputs

Record swapping was used in 2001 Census , in conjunction with SCA in England, Wales and Northern Ireland, while record swapping (without SCA) was employed in Scotland. Swapping has also been used to protect census outputs in the USA. Over-imputation has not been used as an SDC tool and the detailed methodology would have to be developed further if selected as the strategy. The ABS post-tabular method was used in the Australian Census of 2006, and the IACP method is a further development of that, which has not yet been used for a real-life output.

4.3 Origin-destination tables

Origin-destination (O-D) tables are defined in Appendix E, Glossary. There are many difficulties in protecting these tables. Post-tabular methods may provide some protection but have a significant impact on data utility (at low geographical levels) since flows will disappear from the table. For the pre-tabular methods it is likely that highly improbable (but not impossible) flows will occur in the protected table, e.g. cycling or walking 60 miles to work. These issues have been previously discussed at the UK SDC Working Group and the recommendation made that protection for O-D tables (particularly at the low geographical levels) should be provided by licensing. At higher geographic levels an SDC method could be applied or it may be determined that no additional protection (other than aggregation) is required since the flows are less disclosive (this will depend on variable breakdowns).

4.4 Other considerations

Since only geography is perturbed within the record swapping method it is unlikely that inconsistent or illogical records will be created as a result of applying the method. The same is true when only geography is imputed. However imputing other variables will require edit checks to highlight any illogical records, e.g. a 14 year old widow. Non-geographical over-imputation would probably have to be done immediately after, or combined with, the main imputation process which deals with missing values. This would allow over-imputation to take advantage of the post-imputation consistency check and allow consistency in derived variables. Another consideration is that over-imputation will be looking for an 'incorrect' new value while imputation for non-response will go for a 'best'.

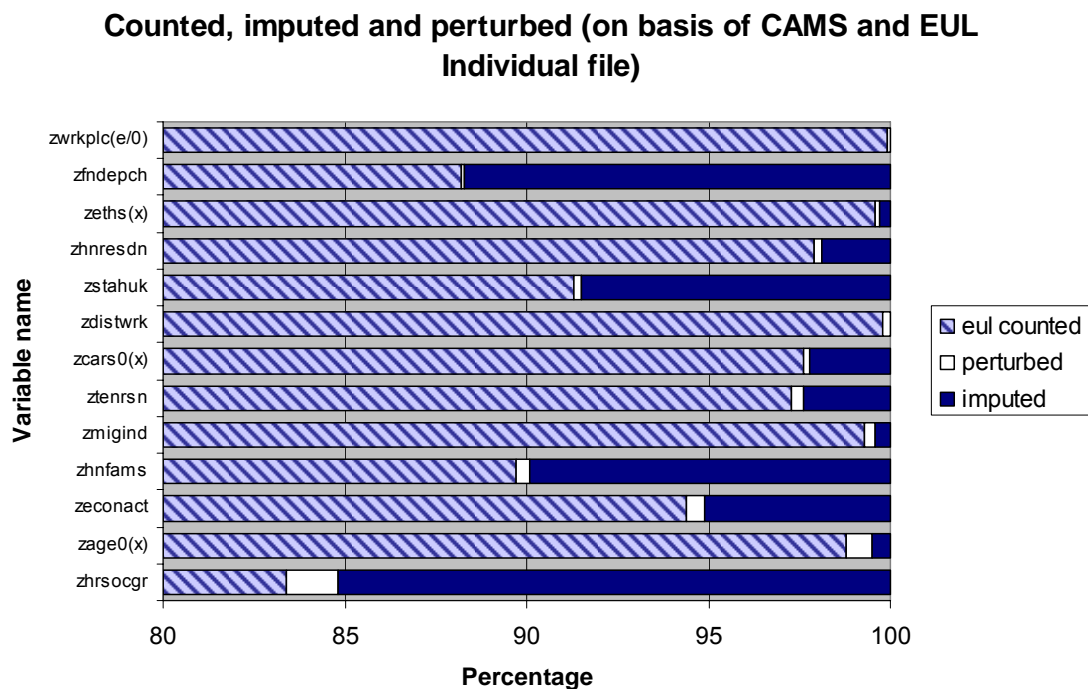
The look-up table used within this analysis for the IACP perturbation method and SCA do not perturb zeros and hence no structural zeros will be altered, thus avoiding illogical cell values. However, not perturbing zeros has a price in terms of managing disclosure risk, see section B2.

4.4.1 Balance between SDC and imputation for unit and item non-response

One argument that has been raised by users is that disclosure control is unnecessary, at least in some areas where response to the census is poor. Where there is low response, persons and households will have been imputed through the equivalent of the 2001 One Number Census. Hence where

subsequently a cell count of one appears in a table, there is less certainty that this corresponds to a 'real' census respondent. There will also be item imputation where an item is missing or is found, subsequent to capture, to be inconsistent with other variables in the individual or household record. The level of item imputation differs considerably between different variables but Wathan (2009) has used the different versions of the 2001 Samples of Anonymised Records (SARs) to examine the level of imputation and compare to the perturbation necessary to move from the 'raw' file available in the CAMS safe-setting⁵ to the non-disclosive version available under end user licence. Figure 4.1 shows that the level of conventional imputation (due to item or unit non-response) was considerably larger than the level of perturbation necessary to produce a non-disclosive file. For instance, the variable 'zhrsocgr'⁶ at the foot of the graph indicates that in the end user licence version, 83.2 per cent of values were as counted (captured), 15.4 per cent had been imputed and 1.4 per cent had been perturbed as part of SDC. This would suggest that the 'light touch' recommended by the Registrars General may be sufficient for protecting tabular outputs, since the data will have already gone through significant amounts of edit and imputation prior to being protected through disclosure control.

Figure 4.1 Balance of counted, imputed and perturbed values in 2001 Samples of Anonymised Records (SARs)



⁵ The Controlled Access Microdata Sample (CAMS)

⁶ Social grade of household reference person

Source: J.Wathan (2009) Imputation and Perturbation in the SARs: A user perspective. Slide copied from presentation at University of Manchester, 30 April.

4.4.2 Record swapping

There were some cases where a match could not be found. This occurred most commonly, but not exclusively, where the household was large. For example, in the 20 per cent swapping, where records were selected randomly for swapping, 19 flagged records (around 1 per cent) could not be matched with another record, even outside the local authority, in the rest of the estimation area; where records were targeted, there were 26 that could not be matched. On the face of it, this is a problem affecting only a small number of records, but on closer analysis, the unmatched records are disproportionately large households and therefore those that are more likely to be risky records. Fewer than half the records flagged in the random swapping that were of size 7 and over were matched and in the targeted swapping none of the 3 records of household size 8 or over were matched. If record swapping were to be used, this would be a key disadvantage.

One solution could be to use record swapping, then to use over-imputation for those records where a match could not be found. This has not been investigated further due to time, but could constitute a compromise between the two methods, combining the strengths of both methods – although this might be a complexity to be avoided.

4.4.3 Over-imputation

If over-imputation were to be selected as the strategy for tabular outputs, the question would remain as to which variables we would perturb in order to protect the data. It might be infeasible to protect all variables and a strategy could be to impute on a limited range of variables to protect a high percentage of tables. The details of which variables were perturbed and the level of imputation would not be revealed to users, and so even those tables that did not contain any of the protected variables would be protected to some extent by the uncertainty created by perception.

Considering tables produced from the 2001 Census can help us here. There are 115 standard tables (for England and Wales) with a total of 52 different variables, having an average of 3.4 variables, ranging from 2 to 5, most commonly 3. As at March 2009, there had been 1,489 tables commissioned from ONS Census Customer Services, Titchfield. These tables commissioned over a period of six years may give a good indication of the types of tables likely to be required or requested in the future.

The variables appearing most in commissioned tables are shown in Table 4.1. The standard tables also have limiting long-term illness, Welsh language and general health in the corresponding top 12 variables.

Table 4.1: Most used variables in commissioned tables

Rank	Variable	Number of tables containing variable	Percentage of tables containing variable
1	Age	685	46.0%
2	Sex	647	43.5%
3	Ethnic group	398	26.7%
4	Qualifications	160	10.7%
5	Occupation	143	9.6%
6	Economic activity	136	9.1%
7	Religion	135	9.1%
8	Country of birth	133	8.9%
9	NS-SEC	123	8.3%
10	Tenure	118	7.9%
11	Household reference person	113	7.6%
12	Industry	106	7.1%

Considering possible variables to be used for non-geographic over-imputation, some may be seen as unsuitable candidates. Sex has only 2 options and in the majority of cases would be imputed back to the true gender. Household reference person is another variable that may have a high proportion being imputed back to the correct person, and would not add much disclosure protection even if it was imputed to a different member of the household. Ethnic group would also be imputed back to the correct value in a high proportion of cases, since it would be imputed on the basis of other variables such as the ethnic group of other household members, including parents, children or other relatives.

When considering the standard tables and commissioned tables combined, a total of 1604 tables, the top nine variables used (excluding sex, ethnic group and household reference person) are:

- a) Age (741)
- b) Qualifications (168)
- c) Occupation (152)
- d) Economic activity (151)
- e) Religion (149)
- f) Tenure (137)
- g) Country of birth (137)
- h) NS-SEC (136)
- i) Industry (112)

These variables cover 1147 of the 1489 commissioned tables and 100 of the 115 standard tables, meaning 78 per cent (1247/1604) of tables have at least one variable, with on average 1.18 of the above variables featuring in each table.

Only 28 per cent of tables have two or more of the listed variables, with just over 8 per cent having three or more. Only 21 tables have four of the above variables and one has five.

In conclusion a large percentage of standard and commissioned tables make use of the same core group of variables, and so choosing these as possible candidates offers the simplest and most effective way of implementing non-geographic over-imputation. The majority of tables can be covered with only a relatively small set of commonly used variables, although adding more variables only extends the coverage by a small amount. If a much higher percentage was required, say 90 per cent of tables having at least one of the variables subject to over-imputation, then several more additional variables would be required.

Further work would be required to find the levels of over-imputation required to provide adequate protection, and on the percentage of tables required to have one or more variables included which are subject to over-imputation.

4.4.4 Additional Rules

Note that over and above the method selected, additional rules will be needed arising from thresholds and from sparsity in tables. Work is currently taking place on differencing between geographies, to ascertain whether it is possible to output on both census (super output areas) and administrative (ward) geographies without increasing the disclosure risk unreasonably.

4.4.5 Protection of workplace tables

Workplace tables may require additional measures from those applied for resident tables. At low geographies, the number⁷ of persons working in an area would be unchanged under record swapping. A previous recommendation is that workplace based statistics could be provided for workplace zones, and residence based statistics for output areas, as long as residence based statistics are not provided for workplace zones (and vice versa). However, the origin-destination tables based on workplace provide a specific challenge, where the origin is residence and the destination is workplace.

It is worth considering too that the data quality of workplace data has been poor in previous censuses. In 2001 UK Census, in England and Wales, a large number [around 40,000 respondents] used non-geographic postcodes for large organisations and many others [7.8 per cent] did not return an address (and therefore a postcode) for their workplace. These are not insurmountable problems and can be addressed post-collection, but they do give an indication as to the data quality of workplace information and therefore

⁷ In addition to the number of persons being unchanged, their characteristics would be unchanged too if workplace is not swapped as well as the corresponding place of enumeration.

create a little uncertainty in cell counts.

4.5 Evaluation Summary

In summary, the short-listed methods have been assessed over a wide range of aspects. The methods were evaluated against a set of evaluation criteria and these criteria have been developed and circulated around the UK Census Offices, and comments and suggested amendments incorporated. There is separate documentation on the development of these criteria, but a grid appears in Appendix D with the weights and scoring.

Even after communication and consultation with the members of the working group, it is clear that the criteria, both on their inclusion and whether mandatory, and their respective weights, are quite subjective. Due to this subjectivity, the decision on SDC strategy was not made solely on the method with the highest score, but the scoring does give an indication as to the overall effectiveness of each method. It is also a useful way of summarizing the large amount of analysis detailed in this report and in highlighting relative differences between methods. Hence the assessment criteria can be seen as just one of a number of tools which contributed to a final decision.

Both record swapping and over-imputation would be able to manage the risk of disclosure and disclosure by differencing. Hence the choice between them may be made on the impact of each method on the data utility.

The weaknesses of over-imputation are that, at the levels of perturbation assessed, the method:

- (i) distorted associations between variables (Criterion S3),
- (ii) impacted on totals and sub-totals within tables at all geographies (though it does not affect the total number of individuals in any geographical area) (Criterion S6)
- (iii) has not been implemented satisfactorily in tests (Criterion S14).

Currently there is no accepted methodology for over-imputation for carrying out disclosure control, which may make it difficult to sell this method to users. Additionally the fact that legitimate data items are removed and replaced with imputed values is likely to be unpopular (Criterion S15). There will also be some outputs, including those at small geographies, where over-imputation would not be applied, since not every variable would be imputed on, e.g. sex, marital status, ethnic group, religion – these would either create difficulties in maintaining consistency with other variables or be very likely to have the real value imputed (Criterion S16).

The weaknesses of record swapping are that

- (i) it could be possible to match high level tables against microdata samples and determine and locate population uniques (Criterion S12)
- (ii) it would be more difficult to protect special populations such as communal establishments and workplaces (Criterion S8).

However, it would be possible to address these issues (i) predominantly through licensing arrangements and (ii) through careful design of the record

swapping methods. It is also more difficult for record swapping to take into account the data quality of different variables (Criterion S4) but it would consider the data quality related to response rates and response-related imputation.

The key strength of record swapping is that no persons or data items are removed from the Census data and therefore outputs at national level and high geographies will be unaffected by record swapping. Record swapping has also been used before (in the UK and USA) to protect census tables, whereas over-imputation has not.

Record swapping was therefore recommended as the primary disclosure control method for 2011 Census. This recommendation was accepted by the ONS Statistics and Policy Committee, and signed off by the UK Census Committee on 25 August 2011. It was agreed that targeted swapping was the preferred method, and the methodology for targeting 'risky' records is being developed for use in 2011.

5 Future work

Further work will be carried out to establish the details of how record swapping will be implemented, as well as other aspects of disclosure control. This will include the following:

- Mechanisms for releasing data, including considerations of hypercubes and flexible table generation
- Levels of perturbation, i.e. percentage of records to be swapped
- Strategies of targeting records to be swapped
- Consideration of swapping rates in areas with high/low imputation due to non-response
- Provision of outputs for overlapping geographies and dealing with slivers
- Consideration of SDC issues related to workplace zones
- Population thresholds
- Level of detail to be made available in outputs

Appendix A- Methods

This appendix contains a detailed description of each of the three short-listed methods and how they were applied to the test data used in the evaluation.

A.1 Record Swapping

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 index¹, 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, workplace and address one year ago (address, OA, LAD, etc). This has the same effect as swapping all other variables and leaving geography fixed.

Note that in the 2001 Census, only address of residence was swapped. This left all statistics based on workplace unchanged, including those on businesses (and the people working in them, apart from their place of residence). However, this did protect the total flows, i.e. the numbers living in OA1 who work in OA2. Employing record swapping where all geographic variables are swapped does not protect the total flows, but does affect the characteristics of those contributing to each flow and the characteristics of persons working in each OA (including occupation and industry variables, which would therefore help in protecting individual businesses).

For the targeted swapping, high risk records (those that contributed to small cells in the set of 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 variables which gave rise to the small cells. 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, where possible. This targeting methodology was based on indicators derived from 2001 Census data. For 2011 a more sophisticated algorithm would be developed.

A.2 Over-imputation

A new method of over-imputation had to be devised for disclosure control of census data using CANCEIS (a specially designed package developed by Statistics Canada to impute missing values arising from item non-response). Imputation in general is a very complex procedure, one reason being the relationships that exist between variables. CANCEIS is based on a nearest neighbour donor approach, and is designed to impute the 'best' possible values, i.e. as close as possible to the true values. Thus for categorical values

such as ethnicity or housing type it is likely that the exact value will be imputed – providing no protection. In the first stage of the SDC analysis, the variables *age* and *geography* were chosen for imputation. Age was chosen 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 that a value close to the original will be imputed, giving some protection but not distorting the data too severely. In addition, *geography* was chosen for imputation since it is commonly associated with disclosure risk as at low levels, it can be used to help identify individual households and persons.

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 LAD⁸ 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 (targeted imputation). 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.

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.

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

In summary, after over-imputation had been applied households which were selected had the variables *ed code* and *ward* imputed but they remained within the same LAD. After over-imputation households which were selected had age imputed: approximately 10 per cent of these ages had exactly the same value imputed back (no change), approximately 45 per cent had an age within one to four years difference from the original imputed, 30 per cent had an age five to ten years difference from the original imputed, the remaining approximately 15 per cent had an age greater than 10 years different from the

⁸ Except where a match might not be found, and a small number of records will take swapped between LADs

original value imputed. CANCEIS aims to minimise the possibility of edit failures (eg. a 10 year old child married to an 80 year old adult).

Subsequent to the comments received from the SDC Working Group and the UKCDMAC SDC sub-group, it was decided to investigate a second type of over-imputation, following some preliminary work that appeared promising. This would involve imputing on the non-geographic variables instead. For the purposes of the evaluation, the two tables studied included the variables country of birth (COB), religion and sex (Table 1) and age, sex and marital status (Table 3). Both used the data from the SJ estimation area (see 3.1). To compare the geographic and non-geographic imputation methods, a 2 per cent perturbation level was used. This involved identifying 2 per cent of records, (i) selected at random and (ii) targeted towards 'risky' records. Within this set of records, over-imputed was performed on one or more of the five variables: country of birth, religion, sex, age and marital status. These were labelled A to E and each of their thirty one possible combinations⁹ was given an equal chance of being flagged for over-imputation, so that, in many selected records, more than one variable was imputed. Thus for one in thirty one of the selected records, all five variables were imputed. The mechanics as to how this might be carried out in a real census situation would require further analysis.

A.3 ABS Cell Perturbation

ABS Cell Perturbation¹⁰ is being used at the ABS for their 2006 Census where it is referred to as 'introduced random error'. The method was developed in response to a need for flexible table generation and also for consistency in generated tables to protect against disclosure by inconsistency.

The ABS method is post-tabular; table cell values are perturbed by values drawn from a 'look-up table'. For each cell in a table, the perturbation in the look-up table is dependent on the original cell value as well as the particular combination of records which were used to compose the cell. Thus the essence of this method is to ensure consistency as the same cell composed of the same records is always perturbed in the same way. This works by assigning a *record key to each* record in the microdata so that, when records are combined for a particular cell using a special function, a *cell key is generated*. The cell key acts as a random value (always the same for the same cell) to draw a perturbation from the distribution of possible 'protected' cell counts, given the unprotected cell count.

Within this framework the method is entirely flexible and the design of the

⁹ The 31 combinations of variables (A-E) are A, B, C, D, E, AB, AC, AD, AE, BC, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE and BCDE.

¹⁰See web reference:

<http://www.abs.gov.au/AUSSTATS/abs@.nsf/vwDictionary/Introduced%20random%20error?opendocument>

look-up table (determining what perturbations are added to what cells) can be specified by the statistical agency according to their requirements. Typically the look-up table would be designed so that perturbations are unbiased, have a defined variance and have limits (to avoid non-negative perturbed cell values). From a simplistic point of view, the ABS method can be described as adding a small, possibly zero, random perturbation to each cell value (hence creating an ‘introduced random error’).

For this evaluation a look-up table was designed so that the frequency distribution of the cell values in the table was approximately preserved; so that approximately the same numbers of ones, twos, threes, etc are in the perturbed table as in the original table¹¹. In other words the perturbation matrix is invariant with respect to the vector of frequencies of each cell value. This is a novel idea designed to further improve the utility of the protected data and thus will be referred to as **IACP (Invariant ABS Cell Perturbation)** from here on. Three look-up tables (that control the perturbation added to the census tables) were devised; these will result in approximate perturbation levels of 2, 10 and 20 per cent. It is difficult to have an entirely consistent comparison between IACP and the pre-tabular methods. The perturbations of 2, 10 and 20 per cent refer to percentage of *cells* perturbed as opposed to the percentage of *records* perturbed for record swapping and over-imputation.

Additivity was restored in all three census tables (after the perturbation stage) using an iterative proportional fitting program. This generally results in a slight loss of consistency. A step by step illustration of the method follows.

1. Each record in the microdata (Table A1) has a record key assigned to it. A record key is a random number between 0 and m. Suppose m = 100.

Table A1. Example of record keys in the microdata for use in IACP method

Record ID	Census Variables					Record Key
	...	Age	Gender	Employment	...	
1	25
2	34
3	98
4	...	41-60	Female	Employed	...	22
5	10
6	55
7	...	41-60	Female	Employed	...	81
8	78

¹¹ Note that because cell counts of zero cannot be perturbed, since there are no record keys to make up the cell key, necessarily there will be more zeros in the protected table – see Tables A2-A3, if cell counts of 1, 2 etc are perturbed to zero.

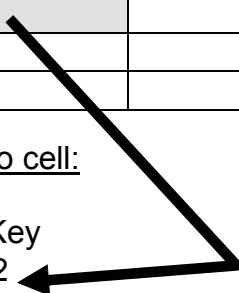
2. For each cell in the table to be protected, a *cell key* has to be calculated. This is done for every cell in the table. See Table A2 for how to derive cell key for a particular cell.

Table A2. Derivation of cell key from record keys in IACP method

Age	Males, employed	Females, employed	Males, unemployed	Females, unemployed
0-20				
21-40				
41-60				
61-80				
80+				

Records contributing to cell:

Record ID	Record Key
Record 4	22
Record 7	81



Cell Key: $[\sum recordkeys] \text{ mod } m$

: where modulus m is the remainder after division by m. This function ensures that the cell keys are also uniformly distributed between 0 and m.

In this example, the cell key would be:

$$81 + 22 = 103$$

$$103 \text{ mod } 100 = 3$$

3. Based on the original cell value and the cell key, the perturbation can be read off from the look-up table (Table A3). This is done for all cells in the table to be protected.

Table A3. Example cell value – Cell key look-up table for IACP method

Example Look-up table		Cell key						
		0	1	2	3	4	5	...(up to m)
Original cell value	0	0	0	0	0	0	0	...
	1	-1	+1	0	0	0	+1	...
	2	-1	0	0	+1	-1	0	...
	3	0	0	0	0	0	-1	...
	4	0	-2	0	+1	0	0	...
	5	-1	0	+1	+2	0	0	...

- Each row acts as a distribution from which the perturbation is drawn at random using the cell key.
- Rows of the look-up table may cycle so that anything with an original cell value of 10+ for example, is drawn from the last n rows of the look-up table (to avoid specifying 1,000 rows or more!)
- The specific values in the final look-up tables would be highly confidential! Part of the disclosure protection is the uncertainty as to where perturbations are added to produce a final cell count.

In our example, the original cell count is 2, the cell key is 3. This corresponds to a perturbation of +1 in the look-up table so that the final protected cell count is $2+1 = 3$.

4. After the perturbation stage, the table doesn't add up, so additivity may need to be restored.

An algorithm such as IPF (iterative proportional fitting) can be applied to the table to make the rows and columns add up. Use of linear programming techniques can optimally minimise additive perturbation to the cells by specifying constraints accordingly. However after the additivity stage, consistency is always lost to a certain extent (unless additive perturbations are all zero).

Look-up table/probability transition matrix

The look-up tables used in this evaluation were set up so that the invariance property is achieved. To do this the look-up table is viewed as a probability transition matrix where the rows represent the original cell values and the columns the values to which the original cell value is changed. So in the example below an original cell value of '2' has a probability of 0.8 of remaining '2' and a probability of 0.2 of increasing or decreasing by one.

Table A4. Example probability matrix for IACP method.

Example probability matrix	Perturbed cell value						
	0	1	2	3	4	5	...

Original cell value	0	1	0	0	0	0	0	...
	1	0.1	0.8	0.1	0	0	0	...
	2	0	0.1	0.8	0.1	0	0	...
	3	0	0	0.1	0.8	0.1	0	...
	4	0	0	0	0.1	0.8	0.1	...
	5	0	0	0	0	0.1	0.8	...

The probability matrix would ideally have the largest weight either on or near the diagonal (representing the probability of no change or a small change). The advantage of viewing the look-up table as a probability matrix is that it is easy to set the perturbation rate. For example, for a perturbation rate of 20 per cent, we achieve this with the IACP method by setting the diagonal to 80 per cent. However, as previously noted this would not be directly equivalent to a 20 per cent swapping or imputation rate since for the pre-tabular methods the perturbation rates apply to records but the perturbation for the IACP method applies to cells in the table.

Moreover, we can set the values of the probability matrix so that the method is invariant with respect to the original frequencies – IACP. This is explained mathematically in Shlomo and Young (2008). In simple terms, the main diagonal and off-diagonals in the probability matrix are adjusted very slightly to balance out the distribution of original cell frequencies in the unprotected, original table. Since for each table to be protected, there is a different distribution of cell frequencies, then the look-up table is adjusted each time for each table. The consequence of this is that the distribution of frequencies is *approximately* preserved in the perturbed tables and the need for further additive perturbations is minimal compared to a ‘standard’ look-up table which considers the rows of the look-up table independently.

Table A5 shows an example row of a look-up table. In order to make use of the cell keys with the probability transition matrix, they are calculated as before, based on a modulus function of the record keys. The cell keys run from 0 to m and are then transformed into a 0 to 1 distribution. If each row of the look-up table is considered separately, with the probabilities thought of as cumulative between 0 and 1 as in Table A5 (because each row of probabilities must sum to one), then the transformed cell key identifies a value on the cumulative distribution which corresponds to a perturbed cell value. Since the transformed cell key is always the same for the same cell, the perturbed cell value is the same, thus achieving consistency.

Table A5. Cumulative probabilities for use in IACP method

Probability matrix	Perturbed cell value						
Original cell value	0	1	2	3	4	5	...
4	0	0	0	0.10	0.85	0.05	...

<i>Cumulative probability</i>	0	0	0	0 - 0.10	> 0.10 - 0.95	> 0.95 - 1	
-------------------------------	---	---	---	----------	---------------	------------	--

A transformed cell key of 0.19 would therefore result in a perturbed cell value of '4'.

Appendix B – Quantitative evaluation

This section of the appendix describes in the detail the approach used for the quantitative evaluation; the data used, the risk measures and the utility measures.

B.1 The data

For EA SJ (Southampton, Eastleigh, Test Valley, 437,744 people, 182,337 households), the following four census tables were analysed, measuring risk and utility by comparing the original and protected tables. The numbers of categories per variable are shown in parentheses:

(Table 1) Country of birth (2 - UK, non-UK) by sex (2) by religion (8) by ward

(Table 2) Number of persons in household (4) by accommodation type (3) by OA / ED

(Table 3) Age (16 age-groups) by Sex (2) by marital status (2 - single, married) by OA / ED

(Table 4) Origin-destination flows from OA / ED to TTWOA where TTWOA is travel to work OA for all in England and Wales.

The set of data used for assessing geographic over-imputation in EA SJ was slightly different to the data used for record swapping and the IACP method (and later the second stage of the evaluation, see 3.1.1). The former, referred to as CPCD data, are partially edited census data which were prepared for use in the development of CANCEIS. The latter, referred to as ORCD data, were raw census data. Table B1 illustrates how these two datasets differ.

Table B1. Differences between CPCD and ORCD datasets

	CPCD (used to carry out over-imputation)	ORCD (used to carry out record swapping and IACP method)
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 per cent 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 <i>address</i> can be used to match geography from ORCD file, before imputation).	All variables available
-------------------	--	-------------------------

For this reason, the census tables assessed in terms of disclosure risk and data utility differed slightly as outlined below:

(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.

In the second stage of the evaluation we followed the recommendation of the working group to consider the two-dimensional sub-groups within Table 1 and Table 3 (see section 3.1.1). Therefore risk and utility were measured for six sub-groups:

- country of birth x sex x ward
- country of birth x religion x ward
- sex x religion x ward
- age x sex x OA
- age x marital status x OA
- sex x marital status x OA

These tables have a level of artificiality, but they were nevertheless valuable for testing the methods. The results are given in Appendix C.

Due to time constraints, Table 2 was not analysed in the second stage.

Despite the differences between the CPCD and ORCD files, the objective was 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 must be considered when interpreting results.

EA KB (Congleton, Chester, Crewe and Nantwich, Ellesmore Port and Vale Royal, 523,465 persons, 215,869 households) is a more rural area in comparison to SJ. Three census tables were analysed with the purpose of assessing more specific features of the SDC methods, particularly additivity, consistency and disclosure by differencing:

(Table 5A) Age (9) by ethnic group (17) by sex (2) for all persons in 133 wards

(Table 5B) Age (9) by ethnic group (17) by sex (2) for persons without limiting long term illness in 133 wards

(Table 5C) Age (9) by ethnic group (17) by sex (2) for all persons with LLTI in 133 wards

Table 5B relates to a subpopulation of table 5A. Table 5C is obtained by differencing table 5B from table 5A. For our evaluation this means that table 5C will not be produced as a disclosure-controlled table independent from tables 5A and 5B but as a derived table via the two disclosure-controlled tables 5A and 5B. This scenario represents a situation common to the 2001 Census where census users made 'special requests' for specific tables which were very similar in composition and differenced tables could be produced indirectly. The variable definition of the tables has also been chosen in order to produce small cells and thus these last three tables allow us to assess the following properties of the SDC methods:

- *Disclosure via differencing* – tables 5A and 5B may be protected but does the (indirectly) derived table 5C have enough protection using the SDC method?
- *Consistency* – the variable definitions for tables 5A-5C are exactly the same although the populations they are based on differ. How well does the SDC method go towards consistency between related table cells?
- *Additivity* – do the table rows and columns add up and are they consistent across the similar tables using the SDC method?

Table B2 provides summary statistics for all the tables considered in this evaluation.

Table B2. Summary statistics for 2001 Census tables used in the evaluation

	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	10,836
Small cells	12%	12%	10%	3%
Zeros	20%	8%	63%	58%
Average cell size	195	249	25	40

	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		

	Table 5A	Table 5B	Table 5C	
Total number of cells	40,698	40,698	40,698	
Small cells	15%	13%	5%	
Zeros	73%	77%	88%	
Average cell size	13	10	2.2	

B.2 Disclosure risk measures

The measurement of disclosure risk is based on the notion of attribute disclosure, i.e. learning something new from the census data about an individual or group of individuals that was not previously known,. This section describes examples of attribute disclosure from tabular outputs and examples involving small counts in cells of tables, and specifies the quantitative measures which will be used to assess the impact of the short-listed SDC methods on disclosure risk. For the purposes of this discussion, there is a table for each geography and with one variable as a row and one variable as a column.

B.2.1 Group disclosure.

Group disclosure occurs when all respondents fall into a single response category for a particular variable for a given category of the other variable. The group of persons with this given category, will then be in the single response category in the first variable. If an intruder knows that a person is in the group, he then learns that the person has the characteristic corresponding to this single response category.

This disclosure is measured within tabular outputs by comparing the percentage of rows / columns with one disclosive non-zero cell in the same location in the original and protected tables. For example, we are measuring the percentage of rows that have not been protected, i.e. the risk left in the protected table. It should be borne in mind that these rows and columns will be protected because the SDC method will create rows and columns that falsely appear to display group disclosure. Although the analysis has not measured the creation of false group disclosure, it may be assumed that this product of the method will be proportionate to the amount of group disclosure removed.

Group Attribute Disclosure (rows):

$$GAD_{rows} = \frac{\sum I_{rows \text{ where ALL respondents fall into same category (X) in O and P tables}}}{\sum I_{rows \text{ where ALL respondents fall into same category (X) in O table}}$$

If no rows or columns exist where all cells are zero except one then the group attribute disclosure measures defined above will equal zero. Where no SDC has been applied the group attribute disclosure measure will equal one.

B.2.2 Within-group disclosure.

This occurs when responses are spread across two categories and one of these categories only contains one person. This person will know that everyone else falls into the other category and will be able to deduce characteristics of the individuals in that group.

This is measured within tabular outputs by comparing the percentage of rows/columns where group disclosure could actually occur in the both the original and protected tables.

Within Group Attribute Disclosure (rows):

$$WGAD_{rows} = \frac{\sum I_{rows \text{ where ALL respondents fall into same 2 categories (X}_i \text{ and X}_j \text{) in O and P tables (only 1 respondent in one)}}{\sum I_{rows \text{ where ALL respondents fall into same 2 categories (X}_i \text{ and X}_j \text{) in O table (only 1 respondent in one)}}$$

If no rows or columns exist where all cells are zero except one then the group attribute disclosure measures defined above will equal zero. Where no SDC has been applied the group attribute disclosure measure will equal one. This measure provides an indication of the level of true within group attribute disclosure remaining in the protected table.

B.2.3 Negative attribute disclosure.

This occurs when no responses fall into a row or column. One can then infer that no one in the population of the table has certain characteristics, determined by the row or column variables.

$$NAD_{rows} = \frac{\sum I_{rows \text{ where no respondents in O and P tables}}}{\sum I_{rows \text{ where no respondents in O table}}}$$

B.2.4 Small cells.

Here the focus is on small cells in tabular outputs, and the following disclosure risk measures can be used to quantify the risk of identity disclosure in tabular outputs:

Percentage of small cells in the table that are not protected:

$$DR = \frac{\sum_{i \in \{1,2\}} I(C_i \text{ unchanged from original value})}{|C_1 \cup C_2|}$$

Where I is the indicator function having a value of 1 if true and 0 if false, C_1 is the set of cells with value 1, and C_2 is the set of cells with value 2, and $|C_1 \cup C_2|$ is the number of small cells in the protected table with the value of 1 or 2.

Percentage of cell counts of 1 in the table that are not protected:

$$DR = \frac{\sum I(C_1 \text{ unchanged from original value})}{|C_1|}$$

B.2.5 Disclosure by differencing.

This occurs when two or more tables taken together enable, by subtraction or deduction, the value for a small cell (1 or 2) to be calculated and the above disclosure risk situations could apply. For example, a table containing the elderly population in private households may be subtracted from a table containing the total elderly population, resulting in a table of the elderly in communal establishments. This table can be quite sparse compared to the two original tables. In the evaluation disclosure by differencing is obtained by assessing disclosure risk (using the measures described above) for Table 5C which is obtained by differencing Table 5A and Table 5B.

B.3 Utility measures

The Information Loss Software¹² was used to evaluate the information loss associated with the short-listed SDC methods. The software calculates a variety of information loss metrics by comparing the protected data with the original pre-disclosure controlled data. The formulae derived here are for the output area geography, but are equally applicable at other geographical levels and the formulae should be adapted as is appropriate:

A) Distortion to distributions as measured by distance metrics

Let D^k represent a table for OA k and let $D^k(c)$ be the cell count for each cell c for OA k . Let $|OA|$ be the number of OA's in the EA. The distance metrics are:

i) Average Absolute Distance (AAD)

¹² Infloss software package has been developed in-house in SDC Methodology.

The Average Absolute Distance is the most intuitive distance metric and measures the average perturbation per cell;

$$AAD(D_{orig}, D_{pert}) = \frac{1}{|OA|} \sum_{k=1}^{|OA|} \frac{\sum_{c \in k} |D_{pert}^k(c) - D_{orig}^k(c)|}{|k|}$$

where $|k| = \sum_c I(c \in k)$ the number of non-zero cells in the k^{th} OA

ii) Relative Absolute Distance (RAD)

$$RAD(D_{orig}, D_{pert}) = \frac{1}{|OA|} \sum_{k=1}^{|OA|} \sum_{c \in k} \frac{|D_{pert}^k(c) - D_{orig}^k(c)|}{D_{orig}^k(c)}$$

The RAD is undefined when the original cell count is zero.

iii) Hellinger's Distance (HD)

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

The Hellinger's Distance Metric is based on Information Theory. It is heavily influenced by small cells and large cells have little impact. A Hellinger's Distance of close to zero is best.

The formulae for the above distance metrics are calculated using the internal cells of the table. The distance metrics can also be calculated using the totals that are aggregated from internal perturbed cells.

B) Analysis on totals and subtotals

The information loss software can perform an analysis of the marginal row totals. These measures can be used to demonstrate how much the additivity has been changed in the table as a result of the protection method.

C) Impact on variance of estimates/ impact on row variance

As for the distance metrics, the variance of the cell counts is calculated at the OA level geography in the table and then an average across all of the OA's is used as the utility measure.

$$\text{Let: } V(D_{orig}) = \frac{1}{|OA|} \sum_{k=1}^{|OA|} \frac{1}{|k|} \sum_{c \in k} (D_{orig}^k(c) - \bar{D}_{orig}^k)^2 \text{ and } V(D_{pert})$$

respectively.

The utility measure is the percent relative difference:

$$RDV(D_{orig}, D_{pert}) = 100 \times \frac{V(D_{orig}) - V(D_{pert})}{V(D_{orig})}$$

The average variance for each row in the original and perturbed table is also calculated by the information loss software along with a confidence interval for

this value.

D) Impact on measures of association based on chi-square tests for independence

The information loss software calculates the Cramer's V statistic before and after SDC has been applied and calculates the percentage difference between these values.

The Pearson Chi-Squared statistic tests if the rows and columns of a table are independent of each other. A low value means that the assumption of independence holds.

The Pearson Chi-Squared statistic is: $\chi^2 = \sum \frac{(O - E)^2}{E}$

where O is the observed frequency and E is the expected (theoretical) frequency asserted by the null hypothesis.

For a $R \times C$ two-way table with counts n_{ij} , $i = 1, \dots, R$, $j = 1, \dots, C$ and

$\sum_i \sum_j n_{ij} = n$, let $\sum_i n_{ij} = n_{i\cdot}$ be the sum of the row and $\sum_j n_{ij} = n_{\cdot j}$ be the sum of

the column. Assuming an underlying multinomial distribution, the expected frequency for the cell under the null hypothesis of independence is:

$e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$ and the Pearson statistic is defined as: $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$.

If the row and column are independent then χ^2 has an asymptotic chi-square distribution with $(R-1) \times (C-1)$ degrees of freedom and for large values the test rejects the null hypothesis in favour of the alternative hypothesis of association.

The Cramer's V statistic is the same as the Pearson Statistic except that it is standardized and used as a "correlation" metric between the rows and columns of the table and obtains a value from 0 to 1. The measure relies on expected cell counts being sufficiently large so as not to inflate the χ^2 value.

Cramer's V is defined as: $CV = \sqrt{\frac{\chi^2 / n}{\min(R-1, C-1)}}$ the associated utility

measure is calculated as the percent relative difference between the statistic for the original and protected table:

$$RCV(D_{orig}, D_{pert}) = 100 \times \frac{CV(D_{orig}) - CV(D_{pert})}{CV(D_{orig})}$$

This illustrates whether the SDC method attenuates the relationship between variables or artificially induces dependencies.

E) Rank test:

A rank test is used to test for changes in the ordering of the cells which can impact on inference with respect to rank correlations. The information loss software only reports the columns in the table where over 10 per cent of the cells have moved groups (The user can decide whether to use 10 or 20 groups per column). In our analysis, the cells are ordered in both the unprotected and protected table separately, and the percentage of cells that have changed decile is measured.

Appendix C – Results

This section of the appendix provides the results of the second stage of the quantitative evaluation of the short-listed SDC methods, along with some results from the first stage (some other results were invalid after errors were found in some versions of the microdata). The first section here provides the results for the evaluation of risk, the next section describes the results of measuring utility.

Record swapping and over-imputation, both geographic and non-geographic, have been assessed at the 2 per cent level, and with records selected both at random and targeted (risky). (In this appendix, “imputation” should be taken to mean “over-imputation” as described in Appendix A, section 2.) Results for both IACP method and for record swapping with small cell adjustment (SCA) method have also been included where possible.

Where targeting methods were assessed, it should be noted that these were based on indicators derived from 2001 Census data. For 2011 a more sophisticated algorithm would be developed. In general the results show that targeted methods give better results than random ones, but this will be much more the case when the new targeting methodology has been developed.

C.1 Risk

Following the recommendation of the working group, see section 3.1.1, we considered the two-dimensional tables within Tables 1 and 3, see B.1.. For each of the six sub-tables, the various types of disclosure were studied. Because Table 1 is at ward level, and both country of birth (COB) and sex have only two levels, the numbers of some types of disclosure in the unprotected data are not particularly large. For Table 3, at output area level, there are, in some calculations, over 5,000 instances of disclosure. In reconstructing the microdata, the unprotected datasets used for the random swapping were slightly different to those used for targeted swapping and hence the numbers of instances of disclosure in the unprotected datasets are marginally different.

C.1.1 Group disclosure

Table C1. Instances of group disclosure removed by protecting table divided by instances in raw table

	Table 1			Table 3		
	COB x relig	COB x sex	Sex x relig	MStat x age	Sex x age	Sex x MStat
Random swapping	0.04	0	0.04	0.01	0.02	0
Random swapping + SCA	0.29	0.28	0.49	0.01	0.03	0.72
Targeted swapping	0.03	0	0.10	0.01	0.02	0
Random imputation non-geog.	0.03	0	0.06	0.01	0.02	0
Targeted imputation non-geog.	0.03	0	0.02	0.01	0.02	0
IACP	0	0	0	0	0	0

The two-dimensional sub-tables of Table 1 (COB x sex x religion) contain fewer instances of group disclosure than the sub-tables of Table 3 (age x sex x marital status). This partly reflects the variables and the number of categories, but mostly it is the lower geography (OA/ED as opposed to ward) in Table 3 that increases the incidence of group disclosure. The 2001 method of record swapping with SCA removes the greatest number of group disclosures. This suggests that around half of the instances of group disclosure in Table 1 are actually columns with all zeros and one small cell of value either '1' or '2'. The level of protection afforded by the other methods at the 2% level is fairly small, though imputation performs a little better than record swapping and both perform slightly better than IACP.

Since the look-up table for the IACP method determines that no zero cells are perturbed, protection for group disclosure will only occur when a non-zero cell (which is likely to have a small value) is perturbed to zero. The look-up table could be modified so that zeros are perturbed, but only so that structural zeros are changed to a non-zero value. On the other hand record swapping and over-imputation can perturb zero and non-zero cells, thus providing more protection.

Note that there are different numbers of instances of group disclosure between the methods in Tables 1 and 3 since the datasets used for imputation are different to those for the other methods.

C.1.2 Within-group disclosure

The results for within group disclosure (see Table C2) are similar to those for group disclosure. Table 1 has fewer instances of within-group disclosure than Table 3 and the protection afforded by all the other methods is limited by the 2 per cent level of perturbation, except for the 2001 Census method of record swapping plus SCA. Overall, the results suggest that record swapping, when targeted to risky records, performs best, but there is some protection afforded by both imputation and IACP.

Table C2. Instances of columns with within-group disclosure removed by protecting table divided by instances in raw table

	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.03	0	0	0.01	0.01	0
Random swapping + SCA	1	1	1	1	1	1
Targeted swapping	0.08	0.18	0.10	0.02	0.01	0
Random non-geog. imputation	0	0	0	0.01	0.01	0
Targeted non-geog. imputation	0.03	0	0.05	0.01	0.02	0
IACP	0	0	0.03	0.01	0.02	0

C.1.3 Negative attribute disclosure

Table C3. Instances of columns with negative attribute disclosure removed by protecting table divided by instances in raw table

	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.02	0	0.02	0.01	0.01	0
Random swapping + SCA	0.02	0	0.02	0	0.01	0
Targeted swapping	0.03	0	0.03	0.03	0.02	0
Random non-geog. imputation	0	0	0	0.01	0.01	0
Targeted non-geog. imputation	0	0	0	0.01	0.01	0
IACP	0	0	0	0	0	0

There are some instances of negative attribute disclosure in Table 1, and Table C3 shows that few of these instances are protected at the 2 per cent level by any of the methods. The results for Table 3 indicate that the level of protection for all methods is low at a 2 per cent level of perturbation, with swapping methods slightly better than the imputation methods. Since cells of size zero are not perturbed in the IACP method, no instances of negative attribute disclosure will be removed, but some new zeros will be created from perturbations of small cells which will add uncertainty. Imputation and record swapping rely on donors and other records matching on some variables, so there is very little effect on these cells of size zero.

In summarizing the results of the first three risk measures, they have suggested that the level of perturbation for all of the methods provides a small amount of protection using any of the methods. Further consideration needs to be given as to whether the level of perturbation constitutes the ‘sufficient uncertainty’ required. In cases where there are fewer zeros, both swapping and imputation offer some protection, though swapping provides considerably more when it is targeted to ‘risky’ records, while IACP leaves all zeros unchanged. However that is not to say that there is no protection afforded by the IACP, since some ambiguity is introduced by some of the zero cells in the protected table arising from perturbations to non-zero cells. Moreover, in all methods, there is uncertainty as to whether a zero cell count in a table is a true zero. Also an important but unmeasured contribution to uncertainty comes from the creation of false cases of apparent disclosure.

Much of the protection in practice would be through the user perception, given that there has been some protection employed, without users being made aware of the full details. In 2001, far more uncertainty was added to the data through the edit and imputation procedures, used to counter item and person non-response and inconsistent answers to census questions, than through disclosure control (see Section 4.4.1 for fuller discussion of this). Hence there is some uncertainty introduced into tables showing apparent disclosure even before disclosure control methods are employed.

Another measure for protecting against attribute and group disclosure is the proportion of zero values changed by the perturbation. With both COB and sex having only two categories, there are limited numbers of zeros at ward level. In fact, none of the methods at the 2 per cent level perturb any of the zeros for that combination of variables (COB x sex) in Table 1. Likewise, marital status (MStat) has only two levels, giving a similar effect for the sex x mstat sub-table of Table 3.

Table C4 shows that record swapping does perturb a larger proportion of zeros where the table has fewer (Table 1), but where the table is sparser and has a greater number of zeros (Table 3) there is little to choose between record swapping and imputation. For COB x Sex (Table 1) and Sex x MStat (Table 3) there are far fewer zeros and though it is possible for small cells to be perturbed to zero, none occurred in the analysis here.

Table C4 Proportion of zeros in the raw table changed in the protected table

	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.03	0	0.03	0.01	0.02	0
Random swapping + SCA	0.02	0	0.02	0.01	0.01	0
Targeted swapping	0.02	0	0.05	0.01	0.02	0
Random non-geog.	0	0	0.01	0.01	0.01	0

imputation							
Targeted non-geog. imputation	0	0	0.01	0.01	0.01	0	
IACP	0	0	0	0	0	0	

C.1.4 Small cells

Table C5. Proportion of cells of value '1' in the raw table removed from the protected table

	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.06	0	0.06	0.05	0.05	0
Random swapping + SCA	1	1	1	1	1	1
Targeted swapping	0.13	0.10	0.09	0.06	0.06	0
Random non-geog. imputation	0.05	0	0.04	0.03	0.02	0
Targeted non-geog. imputation	0.07	0	0.05	0.04	0.03	0
IACP	0		0.02	0.01	0.02	

Table C6. Proportion of cells of value '1' or '2' in the raw table removed from the protected table

	Table 1			Table3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.04	0	0.05	0.05	0.06	0
Random swapping + SCA	1	1	1	1	1	1
Targeted swapping	0.12	0.12	0.14	0.08	0.08	0
Random non-geog.imputation	0.07	0	0.04	0.03	0.04	0
Targeted non-geog.imputation	0.08	0	0.06	0.04	0.05	0
IACP	0.02		0.02	0.03	0.03	

Tables C5 and C6 show how the disclosure risk is reduced (as measured by the proportion of small cells changed) after the SDC methods have been

applied. By definition SCA perturbs all small cells so this risk measure is reduced to zero. For the purposes of this measure, we are mainly looking to see the differences in terms of imputation, swapping and IACP. IACP appears to remove the smallest number of small cells, thus leaving the highest risk, but it is not directly comparable to the other methods since the perturbation levels relate to cells rather than records. For both Table 1 and Table 3, imputation perturbs more small cells than IACP does. Targeted swapping perturbs the largest proportion of small cells. For both swapping and over-imputation the effect of using targeted rather than random methods is that a larger proportion of small cells are protected.

C.1.5 Disclosure by differencing

All methods provide some protection for disclosure by differencing. Pre-tabular methods will protect in the same way that they protect against other forms of disclosure, by perturbing the microdata so that when producing two tables and subtracting, the resultant table will be identical to that produced by interrogating the microdata directly for the 'sliver'. Hence the protection will be equivalent to that given to the records in the sliver. On considering the post-tabular method, IACP, Table 5C has been produced in two ways: (i) by differencing between two tables (5A and 5B) protected independently and (ii) by constructing directly from the microdata and then protecting. This does give rise to a small number of cells where differencing causes apparent negative cell counts to appear (see Section C2.8). For instance, if Table 5A and 5B have true cell counts of 6 and 5 (so the true cell count for the difference is 1), they could be perturbed such that the cell count for 5A is less than that for 5B, say 4 and 5.

C.2 Utility

The in-house Infloss program has been used to generate a number of utility measures, comparing the unprotected tables to the protected tables.

C.2.1 Change in variance

Table C7. Ratio of variances between cells – protected / raw tables

	Table 1	Table 3
--	---------	---------

	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	1.000	1.000	1.000	1.000	1.001	0.980
Random swapping + SCA	0.941	0.929	0.922	0.980	0.965	0.959
Targeted swapping	0.996	0.990	1.001	1.000	1.000	0.991
Random non-geog. imputation	0.987	1.000	0.985	0.999	0.999	0.977
Targeted non-geog. imputation	0.970	0.997	0.985	0.999	0.999	0.994
IACP	1.000		1.000	0.997	0.996	

Table C7 shows the ratio of variance comparing the cells of the original table with the protected table. Ideally the ratio of variance should be around one which implies no change after the SDC method has been applied. After swapping, the ratio hovers around one with some below one and some above. Similar results occur for IACP. However non-geographic imputation always resulted in a ratio below one; the variance consistently decreased. This is because the existing data are used to replace blanked values with imputed values and thus they become much more homogeneous. There is no real difference between the random and targeted approach at the 2% level.

C.2.2 Change in level of association

The Cramer's V test was applied across the whole table (rather than by geography) in Table C8(A) and Figure C8, so for example the level of association between age and marital status across the original table 3 is 0.6107. After random over imputation has been applied this value is 0.6109, a percentage change of 0.04.

Table C8(A). Changes in association (Cramer's V) for evaluated methods and tables.

% change	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0	0	0	0	0	0
Random swapping + SCA						
Targeted swapping	0	0	0	0	0	0
Random non-geog. imputation	-0.35	0.94	8.15	0.04	-0.44	1.75
Targeted non-geog. imputation	-0.62	3.77	8.01	0.06	0.15	1.75
IACP	-0.07		0.13	-0.00	0	

Since record swapping does not alter the totals/sub-totals in each of the categories across the whole table then the measures of association will not change. Over-imputation however does have an impact here.

It should be noted that for many of the variables the level of association is low.

We can also take the counts at ward level (in the case of Table 1) and OA level (Table 3), and assess the changes in association at the lower geographies. Those results are shown in Table C8(B). Record swapping generally still performs better than over-imputation though there is some effect on associations at the lower geographies, since records are swapped between wards and OAs, the geographies at which the associations are being assessed. IACP has the least effect on associations while the 2001 method of record swapping plus small cell adjustment has greatest effect, this being most marked where the tables are sparser.

Table C8(B). Changes in association (Cramer's V) for evaluated methods and tables.

% change	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.15	-0.93	0.17	-0.03	-0.15	-0.12
Random swapping + SCA	1.04	-0.93	2.01	1.29	6.95	-0.12
Targeted swapping	0.04	-2.74	-0.35	0.00	0.19	0.06
Random non-geog. imputation	0.89	0.37	5.26	-0.01	0.01	0.85
Targeted non-geog. imputation	0.26	0.59	5.15	0.04	0.26	0.43
IACP	-0.05		0.16	-0.00	0.00	0.02

C.2.3 Hellinger's' Distance

For many of the two-dimensional tables, the results are mixed for the Hellinger's distance metric. The metric shows the difference between the protected and unprotected table so a value close to zero should be best. It may be difficult to compare the methods directly since the results depend on the quality of the CANCEIS imputation on the variables (rather than imputing geography). Table C9 shows that IACP performs best but there are mixed results for the imputation and swapping methods. The over-imputation method appears to out-perform swapping where the variables are straightforward (COB, age and sex) but where there may be greater complexity in the imputation (religion and marital status) there is likely to be greater damage caused by imputation than swapping – unless we adjusted the perturbation rates for individual variable to compensate.

Table C9. Changes in Hellinger's Distance metric for evaluated methods and tables.

	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.333	0.079	0.302	0.319	0.287	0.027
Random swapping + SCA	0.845	0.177	0.893	1.331	1.208	0.036
Targeted swapping	0.432	0.113	0.397	0.357	0.332	0.029
Random non-geog. imputation	1.545	0.070	1.547	0.217	0.223	0.037
Targeted non-geog. imputation	1.572	0.087	1.576	0.240	0.255	0.040
IACP	0.057		0.081	0.202	0.203	

C.2.4 Relative Absolute Deviation

The Relative Absolute Deviation (RAD) values in Table C10 are expressed in percentage terms, so that, for example, on average, across the COB * religion cells in Table 1 the values are changing by 0.36 per cent. The largest percentage changes are incurred by using the SCA method, where small cells will be adjusted, the relative change perhaps a doubling or trebling of the cell count. IACP performs best here, having the lowest relative change, better than both the pre-tabular methods, whether random or targeted. Generally, imputation performs slightly better than swapping except in those tables with the religion variable.

The patterns are similar to those for Table C9 (Hellinger's' distance).

Table C10. Relative Absolute Deviation (RAD) for evaluated methods and tables.

	Table 1			TableE 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	0.360	0.021	0.349	0.712	0.676	0.019
Random swapping + SCA	2.270	0.321	2.233	5.158	4.389	0.046
Targeted swapping	0.710	0.051	0.605	0.850	0.840	0.021
Random non-geog. imputation	0.854	0.017	0.850	0.405	0.494	0.020
Targeted non-geog. imputation	0.900	0.022	0.902	0.492	0.600	0.022
IACP	0.060		0.116	0.375	0.376	

C.2.5 Average Absolute Deviation

Table C11. Average Absolute Deviation (AAD) for evaluated methods and tables.

	Table 1			Table 3		
	COB x Relig	COB x Sex	Sex x Relig	MStat x Age	Sex x Age	Sex x MStat
Random swapping	1.630	2.186	1.527	0.161	0.149	0.305
Random swapping + SCA	1.784	2.275	1.686	0.342	0.301	0.313
Targeted swapping	1.718	2.514	1.552	0.166	0.158	0.301
Random non-geog. imputation	12.46 8	2.500	12.44 8	0.096	0.123	0.348
Targeted non-geog. imputation	12.80 4	2.686	12.71 1	0.106	0.134	0.371
IACP	0.130		0.184	0.082	0.082	

Table C11 considers the absolute rather than relative deviations, and this shows similar findings.

C.2.6 Impact on totals and subtotals

Table C12(A). Impact on totals and subtotals for evaluated methods and tables – Absolute Differences, aggregated across categories for the table as a whole

	Table 1			Table 3		
	COB	Sex	Relig	MStat	Sex	Age
Random swapping	0	0	0	0	0	0
Random swapping + SCA	9	14	14	94	58	43
Targeted swapping	0	0	0	0	0	0
Random non-geog. imputation	98	54	1,720	129	52	20
Targeted non-geog. imputation	246	49	1,757	182	49	33
IACP	5	4	23	46	43	22

Table C12(A) shows the impact on the variable sub-totals across all geographies in each of the Tables analysed. Because every swapped record is still within the dataset, the effect on the totals in each category of the

variables is zero, though there is some effect from small cell adjustment. Imputation impacts far more. Large values for religion imputation may be reflected by the fact that in 2001 religion was not imputed by CANCEIS and there was not a readily available algorithm and set of matching variables. Nevertheless, the results are sufficient to show there is some effect of over-imputation on sub-totals and totals in tables. IACP performs better than over-imputation on Table 1, except for the sex variable, where CANCEIS may be able to 'predict' sex well from other matching variables. Where Table 3 is that much sparser, IACP has greater effect on variable totals. Table C12(B) shows the impact where the variable sub-totals have been calculated for each area and then summed across all wards (in the case of Table 1) and OAs (in the case of table 3). It shows that swapping has less effect than over-imputation. IACP performs best since it affects 2 per cent of cells rather than 2 per cent of records.

Table C12(B). Impact on totals and subtotals for evaluated methods and tables – Absolute Differences, aggregated across categories for each area, summed

	Table 1			Table 3		
	COB	Sex	Relig	MStat	Sex	Age
Random swapping	232	0	1,840	1,400	0	5,616
Random swapping + SCA	401	40	3,327	2,953	1,027	6,929
Targeted swapping	260	0	1,996	1,512	0	6,432
Random non-geog. imputation	770	1,062	4,484	2,086	960	7,876
Targeted non-geog. imputation	1,028	1,162	4,606	2,554	933	8,044
IACP	60	89	19	364	364	45

C.2.7 Rankings

Table C13 shows the effect on the ranking of cells after protection. In our analysis, all cells are ordered in both the unprotected and protected table separately, and the percentage of cells that have changed decile is measured. These are shown for the different methods in Table C13. IACP has least effect and over-imputation is better than swapping where the variables may be more easily imputed to some degree of accuracy – sex and country of birth (two categories) and, to a lesser extent, marital status and age. Where religion is imputed, there appears to be a greater effect on the rankings and this may be due to the greater difficulty in imputing that variable from others in the microdata.

Table C13. Percentage of cells changing deciles when ranked by cell size, for evaluated methods and tables.

	Table 1			Table 3		
	COB	COB	Sex x	MStat x	Sex x	Sex x

	x Relig	x Sex	Relig	Age	Age	MStat
Random swapping	4.2	8.6	5.4	6.8	6.8	3.5
Random swapping + SCA	6.6	2.7	7.5	11.5	10.6	11.5
Targeted swapping	5.2	6.4	4.6	7.9	6.7	3.4
Random non-geog. imputation	9.6	0.7	9.5	4.4	5.3	3.9
Targeted non-geog. imputation	8.8	1.4	9.6	5.2	6.3	4.6
IACP	0.5		1.3	3.0	3.3	

C.2.8 Consistency

Since record swapping and over-imputation are pre-tabular methods and involve perturbation of the microdata before tables are created these methods will always produce consistent tables, i.e. the same cell in a different table will always have the same value. SCA involves randomly adjusting small cells upwards or downwards based on a probability scheme hence the same cell may be perturbed upwards in one table and downwards in another, thus consistency is not preserved. The use of microdata keys in the IACP method ensures that the same cell has the same perturbation each time it falls in a table, however some of this consistency is lost when the table is made additive. Differencing tables protected using the IACP method may also produce inconsistencies.

This was tested using Tables 5A-C. Two versions of the differenced table were produced and tested: Table 5C was derived by using IACP to protect Table 5A and Table 5B and then taking the difference between them; Table 5C* was derived by taking the difference between unprotected Tables 5A and 5B, and then applying IACP to the new table. For this work, we looked at three levels: 98%, 90% and 80% (equating to approximately 2, 10 and 20 per cent of cells perturbed). Tables C14 and C15 summarise the differences between Table 5C and 5C*, and hence the level of inconsistency produced by the IACP method.

Table C14. Zero, non-zero and negative cell counts in differenced table 5C and 5C*

	No. zero cells	No. non-zero cells	No. negative cells
Table 5C* (IACP 98%)	35,960	5,178	-
Table 5C (IACP 98%)	35,842	5,296	76
Table 5C* (IACP 90%)	36,011	5,127	-
Table 5C (IACP 90%)	35,753	5,385	207

90%)			
Table 5C* (IACP 80%)	36,044	5,094	-
Table 5C (IACP 80%)	35,603	5,535	357

Table C15. Number (and percentage) of cell counts differing between Tables 5C and Table 5C*

	Difference between Table 5C and 5C*			
	Total no. differing cells	+/-1	+/-2	>2
IACP 98%	468 (1.2%)	426 (1.1%)	31 (0.08%)	11 (0.03%)
IACP 90%	1,555 (3.8%)	758 (1.8%)	725 (1.8%)	72 (0.2%)
IACP 80%	2,443 (5.9%)	1,871 (4.5%)	383 (0.9%)	189 (0.5%)

The results in Tables C14 and C15 show that the IACP method does not maintain consistency between differenced tables. For example for IACP 98 % there are 468 cells whose values differ between the two tables, the majority of these differ by an absolute value of 1. The maximum difference is 5. For IACP 90% we observe more inconsistencies, the maximum difference in any cell value between Table 5C and Table 5C* is 7. For IACP 80% there are more inconsistencies and the maximum difference in any cell value is 7.

We also considered the row and column totals of Tables 5C and 5C*. For IACP 98% all the row and column totals between the two tables were the same. For IACP 90% and IACP 80% one row and one column total were different in Table 5C* compared to Table 5C; both had an absolute difference equal to one.

C.2.10 Additivity

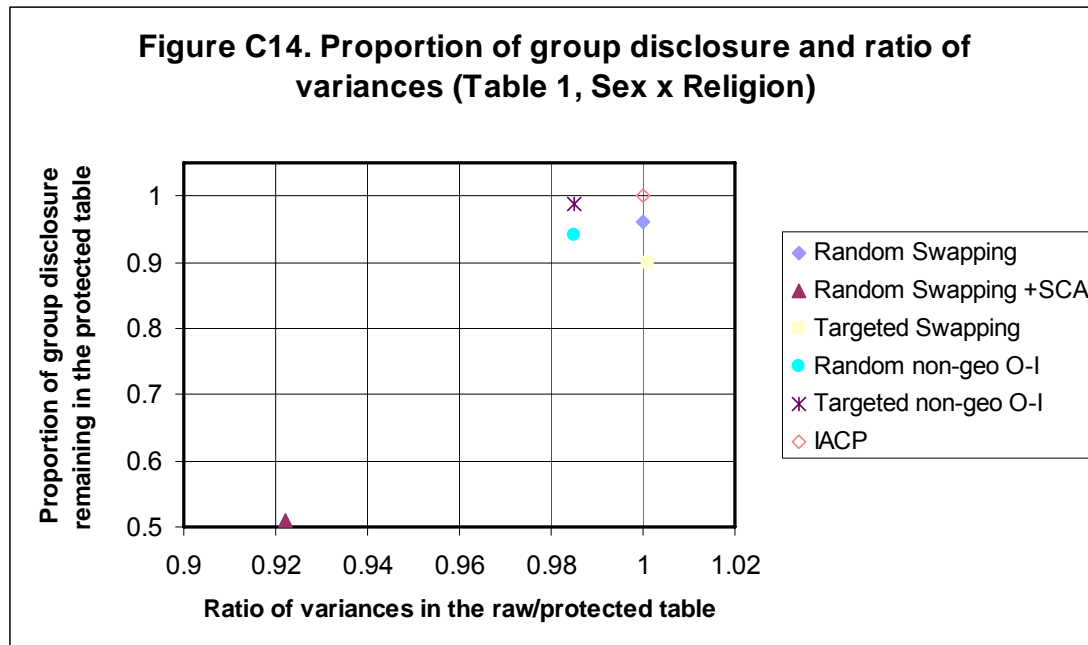
Since record swapping and over-imputation are pre-tabular methods and involve perturbation of the microdata rather than the tables themselves, these two methods necessarily preserve additivity in all tables. The IACP method is post-tabular and involves perturbation of table cells, so additivity is not preserved in all cases. However, the IACP algorithm restores the additivity. Hence all three short-listed methods would produce outputs that are additive.

C.3 Disclosure risk and utility

The following graphs (Figures C14-C19) use the proportion of group disclosure remaining in the protected tables as a measure of disclosure risk, and this is compared to several utility measures. Ideally, we would like to

compare the utility measures across the different methods where the risk is approximately the same.

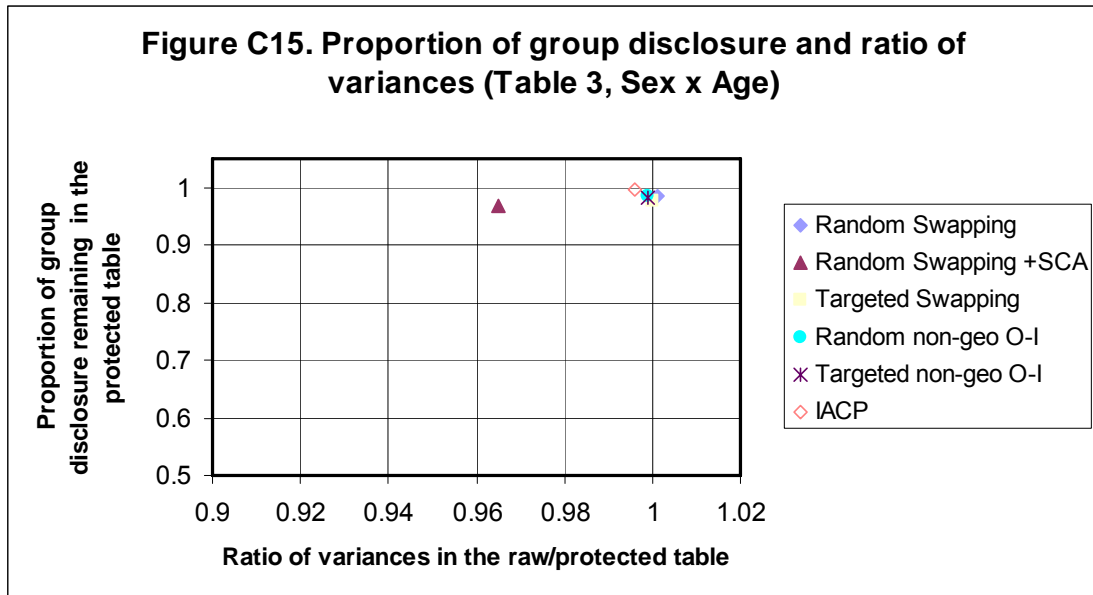
Figure C14. Proportion of group disclosure remaining in the protected table and the ratio of the variance change between the raw and protected tables.



The 2001 method of random swapping with SCA has the largest effect on the variance (reducing variation in protected tables) – see Figure C14. But this method also reduced the group disclosure in the protected tables the most. The other evaluated methods slightly reduced the variances, and this was more noticeable with over-imputation methods. Random and targeted swapping had similar levels of group disclosure remaining in the tables, performing slightly better than over-imputation.

The equivalent comparison using Table 3 data in Figure C15 highlights the similarity in the proposed 2011 methods. Random swapping with SCA has a slightly lower variance compared to the other methods but we are not considering that 2001 method for 2011, only as a comparator with the short-listed methods. There is very little difference in performance between the short-listed methods here.

Figure C15. Proportion of group disclosure remaining in the protected table and the ratio of the variance change between the raw and protected tables.



Both forms of over-imputation perform worst in the Hellinger’s distance metric (see Figure C16). Targeted swapping performs quite well protecting more cells, with similar utility to random swapping. The IACP method leaves the most instances of group disclosure in the protected data but it has the best Hellinger’s score.

Figure C16. Proportion of group disclosure remaining in the protected table and the Hellinger’s’ distance metric showing the difference between the protected and unprotected tables (values closest to zero are best).

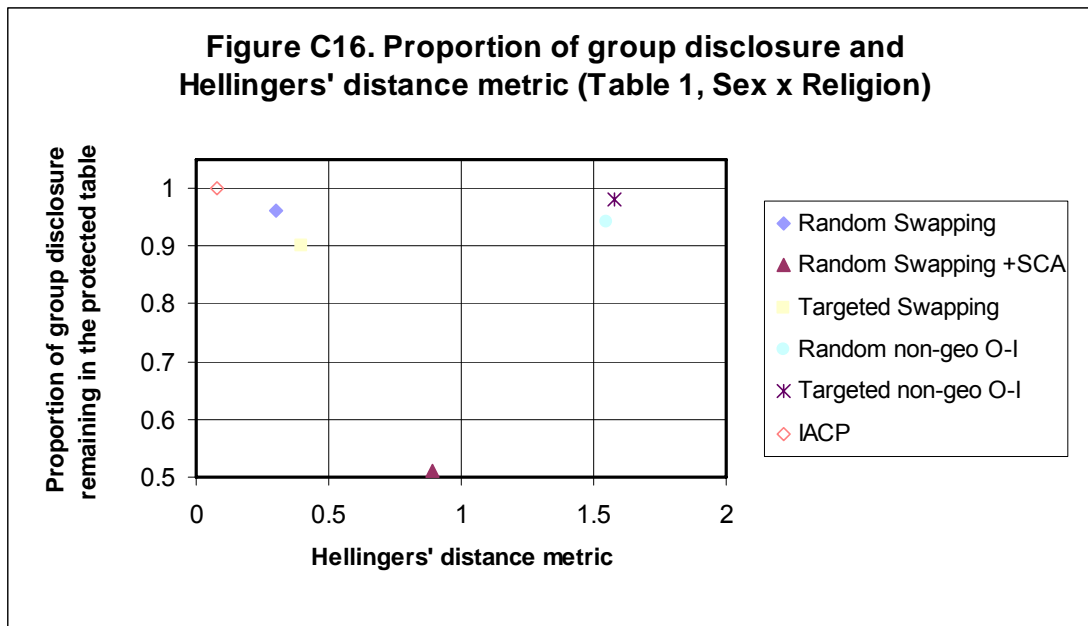


Figure C17. Proportion of group disclosure remaining in the protected table and the Hellinger's distance metric showing the difference between the protected and unprotected tables (values closest to zero are best).

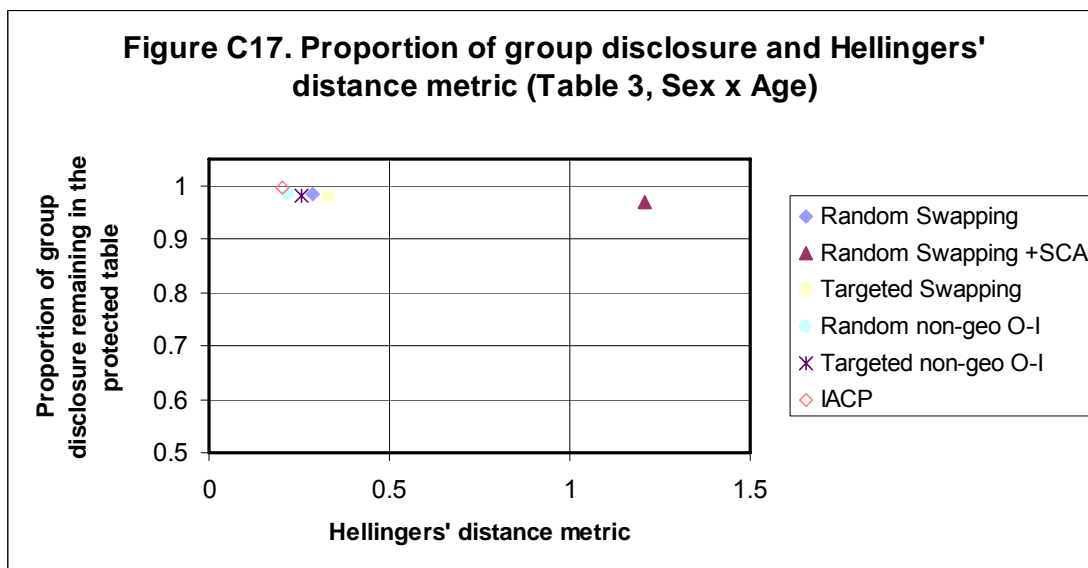


Table 3 data shows the over-imputation methods performing similarly to the record swapping and IACP methods. The 2001 method of random swapping and SCA performs worse on Hellinger's method. There is very little difference in performance between the short-listed methods.

Figure C18. The proportion of group disclosure remaining in the protected table and percentage of cells changing deciles between the raw and protected tables.

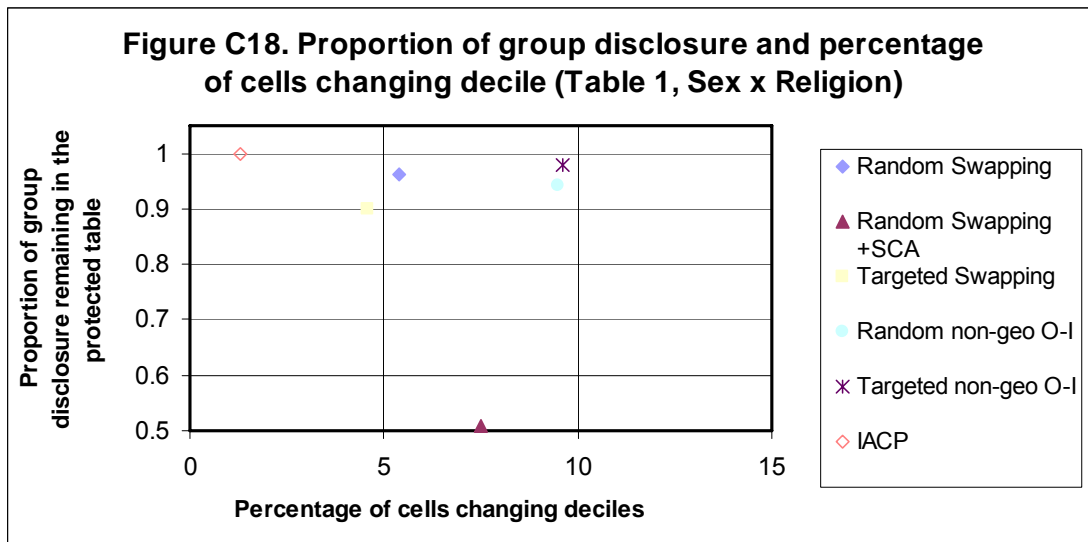


Figure C18 shows the two over-imputation methods performing similarly with a high proportion of group disclosure remaining but a relatively high proportion of cells changing decile. The random and targeted swapping methods perform better with slightly greater protection and better utility on this measure. IACP performs best but has the least protection against group disclosure.

Figure C19. The proportion of group disclosure remaining in the protected table and percentage of cells changing deciles between the raw and protected tables.

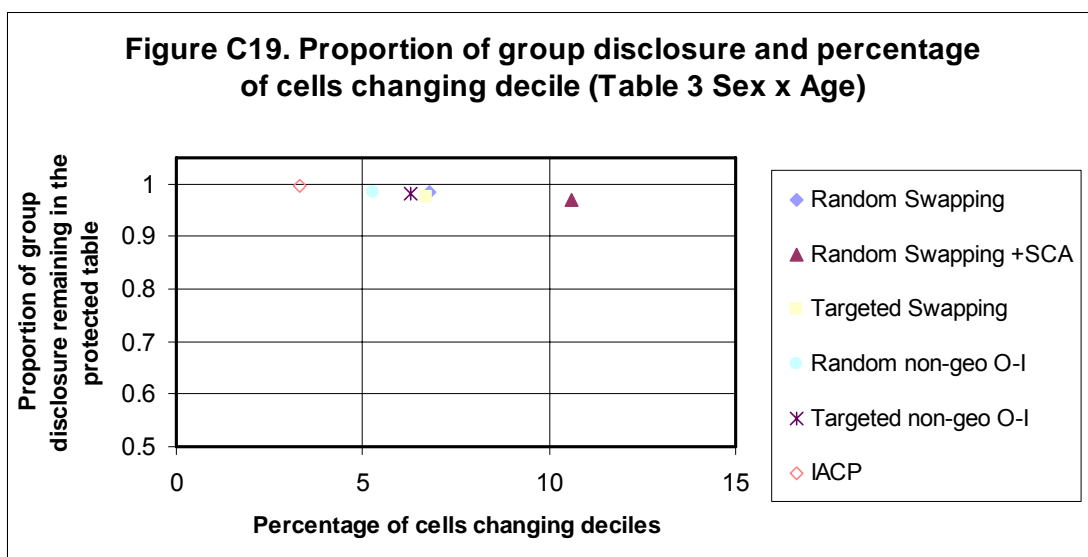


Figure C19 shows a similar analysis for Table 3, in terms of percentage of cells changing deciles. Random swapping with SCA (2001) performs worst and IACP best. There is little difference between the other short-listed methods across a similar level of risk but the over-imputation methods do perform slightly better than the record swapping methods.

C.4 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 per cent of the table cells at OA level, with small cell values making up the majority of non-zero cells, although zero rows are usually suppressed from output (summary statistics are shown in Table B2).

Determining an appropriate SDC strategy for O/D tables is very problematic due to their sparsity. 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 we mean the total numbers of flows between origin and travel-to-work destination (and not the variable breakdown, eg. breakdown of flows into numbers travelling by bike, bus, etc).

Due to the sparsity of the table, limited risk and utility measures are considered. Results are only displayed for 20 per cent swapping and imputation. Here we impute on geography, rather than the non-geographic variables, to attempt to increase the protection, since non-geographic imputation would not change any flows, only some of their characteristics. Due to some errors found later in the microdata, the results for these are approximate and have been rounded to the nearest whole percentage point to reflect this. However, they do give an indication as to the scale of the problem. Results for the IACP method are inferred from the other results.

C4.1 Disclosure risk and utility

Group, negative attribute and within-group disclosure risks arise 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 percentages of non-zero cells and of cells of value '1' which were unperturbed.

Table C16. Percentage of (i) non-zero cell counts and (ii) cell counts of '1', that are unperturbed in origin-destination table, Table 4

	% cells unperturbed (that were not originally zero)	% '1's unperturbed
Random swapping 20%	100%	100%
Random swapping 20% with SCA	1%	0%
Targeted swapping 20%	100%	100%
Random imputation 20%	70%	80%
Targeted ilmputation 20%	67%	77%

O/D tables are extremely sparse so many of the 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. 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.

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

Random swapping with SCA results in only around 1% of total flows being unperturbed. This is because many of the total flows are either '1's or '2's which are small cell adjusted. This is a key problem with SCA as flows are actually 'disappearing'. Random swapping with SCA results in 57 per cent of the perturbed cells (total flows only) having an absolute difference of one but 39 per cent of the perturbed cells having an absolute difference of two or three. This is due to SCA modifying small cell counts on top of swapping. Therefore this approach results in the greatest distortion. 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 around 1 per cent. As expected the percentage of '1's being unperturbed is zero.

Table C17. Absolute Differences for cells in origin-destination table, Table 4

	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%

Over-imputation results in 67-70 per cent of the totals flows being unperturbed for a 20 per cent 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:

Table C18. Example of geographic over-imputation in O-D tables

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. Imputation results in 73-91 per cent of the perturbed cells (total flows only) having an absolute difference of one. 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. Imputation adds uncertainty so a flow of 1 may have been imputed but imputation may lead to inconsistencies when flows are displayed with variable breakdowns, eg. if in the above example the distance from A to Z is great or it is not possible to travel by bus from location W to B.

Swapping of geography (**ie. 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 households are swapped (origins) as well as work locations (destinations). Thus all flows still remain intact but the characteristics of the households making those flows are changed; eg. whether a flow from X to A involves travel by bike or bus. See Table C19 for example.

Table C19. Example of record swapping

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

1		42	location Y		bike	location B
---	--	----	------------	--	------	------------

Persons 1 and 2 have swapped location but the flows between origin and work location remain. Swapping alone results in no change to the perturbed cells (total flows only) as described above for disclosure risk. In terms of totals (column 1 of the O/D table), swapping leaves these unaffected ie. the percentage of cells unperturbed is 100 per cent, because although the households are swapped, the flow is still there (just the swapped households have different characteristics). Swapping is likely to lead to apparent 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 is neither a college nor a university. The same will hold for non-geographic imputation; no protection will be provided for flows, since neither origin nor destination will be perturbed, though characteristics may be imputed.

The IACP method will perturb 20 per cent of non-zero cells, the amount of perturbation will be determined by the look up table. Zeros will not be perturbed, as, with SCA, flows will disappear from the tables when for example a '1' has a '-1' perturbation.

There are many difficulties in protecting origin-destination tables. Post-tabular methods may provide some protection but have a significant impact on data utility (at low geographical levels) since flows will disappear from the table. For the pre-tabular methods it is likely that illogical flows will occur in the protected table, i.e. cycling or walking 60 miles to work. These issues have been previously discussed at the UK SDC Working Group and the recommendation made that protection for O-D tables (particularly at the low geographical levels) should be provided by licensing and restricted access. At higher levels an SDC method could be applied or it may be determined that no additional protection (other than aggregation) is required since the flows are not so disclosive (this will depend on variable breakdowns).

Appendix D. Assessment criteria

Notes:

Shading indicates new criteria suggested by the UKCDMAC SDC sub-group on peer review. Note that these are criteria additional to those agreed by the UK SDC Working Group

The criteria should be scored on a scale of 0-5. Criteria marked 'Mandatory' must receive a score of at least 4 for the method to be considered.

0: The criterion is not met at all

1: The criterion is partly met, but only to a very limited degree

2: The criterion is sometimes met, or to some degree

3: The criterion is usually met

4: The criterion is nearly always met, or almost completely met

5: The criterion is always met

The total score for a method is calculated from Σ (Weighting x Score).

Criterion M1 RS was agreed as scoring either a 4 or a 5 and for S6 OI was agreed as scoring either a 3 or a 4. Both have been given the higher score in the table above.

	Assessment criteria	Weighting	Record swapping		O-I (non-geographic variables)		IACP	
			Score	Weighted Score	Score	Weighted Score	Score	Weighted Score
	MANDATORY CRITERIA							
M1	The method creates the desired level of doubt about any attribute disclosure and protects against differencing	10	5	50	5	50	4	40
M2	Marginal totals in protected tables are unbiased	10	5	50	5	50	5	50
M3	Protected tables are additive	10	5	50	5	50	5	50
M4	The method cannot be unpicked	10	5	50	5	50	4	40
	SECONDARY CRITERIA			200		200		180
S1	Method provides consistent cell counts and totals between different protected tables	9	5	45	5	45	3	27
S2	The method is practical bearing in mind the resources available in terms of manpower, computing power and software costs	8	4	32	4	32	3	24
S3	For a given level of risk relationships between variables are maintained in protected tables	7	4	28	3	21	3	21

	Assessment criteria	Weighting	Record swapping		O-I (non-geographic variables)		IACP	
			Score	Weighted Score	Score	Weighted Score	Score	Weighted Score
S4	The method can take into account the levels of imputation and overall data quality of different variables	6	3	18	5	30	0	0
S5	Counts of households and residents for small areas are not unduly perturbed	6	5	30	5	30	4	24
S6	The method does not unduly perturb/affect counts for large geographies (e.g. LA level and above)	6	5	30	4	24	3	18
S7	The method has a low impact on the variance of estimates	6	5	30	4	24	4	24
S8	The method can be used or adapted to protect outputs from special populations such as communal establishments or from workplaces	6	3	18	5	30	5	30
S9	Will not restrict the detail of releases or the subsequent protection method to be used for microdata samples	6	4	24	5	30	5	30
S10	The method and any required software will have adequate lifespan for purpose	6	5	30	5	30	3	18
S11	The method can easily be accounted for by users in analysis	5	2	10	2	10	2	10

	Assessment criteria	Weighting	Record swapping		O-I (non-geographic variables)		IACP	
			Score	Weighted Score	Score	Weighted Score	Score	Weighted Score
S12	The same method can be applied to microdata outputs	5	0	0	4	20	0	0
S13	The method is likely to be easily understood by users	5	5	25	4	20	2	10
S14	The method has been effectively used for protecting similar outputs	4	5	20	3	12	3	12
S15	The method makes use of all data collected in the Census	7	5	35	1	7	4	28
S16	The method will be applied systematically to all tables and all cells	7	4	28	3	21	5	35
				403		386		311
	TOTAL WEIGHTED SCORE			603		586		491

Appendix E - Glossary and abbreviations

Here is a brief list of some of the abbreviations and terms used in this paper.

ONS - the Office for National Statistics
GROS - General Register Office, Scotland
WAG - Welsh Assembly Government
NISRA - Northern Ireland Statistics and Research Agency
ABS - Australian Bureau of Statistics

SDC - statistical disclosure control; the branch of ONS Methodology which deals with this.

IACP - invariant ABS cell perturbation - a type of post-tabular SDC method using record keys to help maintain consistency between cell values in different tables

CANCEIS - edit and imputation software developed by Statistics Canada

UKCDMAC - United Kingdom Census Design and Methodology Advisory Committee. This is made up of statisticians and academics.
UKCDMAC SDC subgroup / UK SDC subgroup - subgroup of the above which deals with SDC matters relating to the Census.

UKCC - UK Census Committee. This consists of the National Statistician, a representative of the Welsh Assembly Government and the Registrars General of Scotland and Northern Ireland, along with other senior management of the UK Census offices.

Working Group / UK SDC Working Group. This includes ONS staff and representatives from GROS, NISRA and WAG, dealing with SDC working level issues.

Small cells - values in a table which are below an agreed safety threshold.

LA / LAD - local authority / local authority district. Local authorities lie in a two-tier system within the county. In this paper this LA is taken to include unitary authorities.

UA - unitary authority - a type of local authority which has a single tier and is responsible for all local government functions within its area (in that respect it is similar to a county without any underlying districts).

OA - output area. This is the smallest area for which census tables are published.

ED - enumeration district. This is an area defined for data collection.

Further information about geographical terms used in connection with the Census can be found at

http://www.statistics.gov.uk/geography/census_geog.asp

Origin-destination tables - this term is used for two sets of statistics:

(1) Changes in area of residence between the year before the Census and the day of the Census

(2) Journeys between residence and work-place (sometimes termed "workplace tables").

SDC concerns are mainly with the second set.

References

N.Shlomo and C.Young (2008) *Invariant Post-tabular Protection of Census Frequency Counts*. Paper presented at Privacy in Statistical Databases Conference, Istanbul, 24-26 September.

ONS (2006) <http://www.ons.gov.uk/census/2011-census/produce-deliver-data/confidentiality/index.html>

ONS (2007) <http://www.ons.gov.uk/census/2011-census/produce-deliver-data/regrs-gen-agreement.pdf>

ONS (2008) – UK SDC Working Paper, October 2008, evaluation of short-listed methods

J.Wathan (2009) Imputation and Perturbation in the SARs: A user perspective. Presentation at University of Manchester, 30 April 2009.