Responses to UKCDMAC SDC subgroup comments on interim report

This document addresses the main points made by the UKCDMAC SDC subgroup in May and early June, as responses to the interim quantitative evaluation of SDC methods for the 2011 Census. *Sub-group members' comments are italicised here,* while responses to these are in blue.

In our responses we have had to take into account the resources available to carry out this work and the requirement to make a recommendation on the SDC methods for the 2011 Census tables by the end of September. Comments made here have also been reflected in the work plan, separate to this document.

The recommendation to drop over-imputation from the short-list raised some questions, many related to the amount of evidence gathered at the time of the interim report.

The work has only been carried out on a single EA and should be replicated on at least one and preferably several other areas with different demographic properties.
All the results are for one EA, what about other EAs?
There is evidence that the risk profile varies with area, this may interact with the SDC method.

At the time of the interim report, all analysis had been carried out on three tables (plus the Origin-Destination table) in one estimation area, SJ, containing districts in the Southampton and surrounding area. We are commencing work on a second EA, within South Cheshire, which is a more rural area than SJ and considering two new tables that also could be used to look at risk of disclosure by differencing.

There were also several comments on the methodology of over-imputation and the fairness of the comparison with record swapping.

The comparison is unfair because over-imputation perturbs two variables where record swapping perturbs one. Utility would be improved for over-imputation is only imputed one variable.
Imputation could be controlled more by 'hard matching' for the same auxiliary variables as used for record swapping.
I would like to see the effects of just a single variable change on over-imputation.

Age and geography were over-imputed whereas record swapping only modifies the geography. Two of the three tables (see below) did not contain the age variable, meaning that both record swapping and over-imputation were only perturbing geography.

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

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

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

For the two additional tables, we are only imputing geography, so the comparison between the different methods should be a fair one.

CANCEIS only implements one method of over-imputation which is not the only possibility.
Record swapping would need to be redesigned and redeveloped for the 2011 Census whereas software is available for over-imputation.

Both these points are true and are noted. However, we are short of resource to carry out over-imputation via a number of different packages in the timescale available. We are aware that implementing record swapping, and indeed a method similar to that used by ABS, would necessitate some software development if part of the strategy, this is not a criterion on which we would dismiss a prospective SDC method for 2011.

Some comments were received that requested presentation of further analysis.

More analysis on over-imputation would be good; number of recipients where no donor was found, distribution of distances between donor and recipient, how many donors were used multiple times, etc. The reduction in disclosure risk may not necessarily be achieved by swapping a larger percentage of records - this is hypothetical.

These are valid points. Depending on time, we may produce some additional analysis on over-imputation along these lines for the final report.

The choice of the four tables used for testing the methods seems to be appropriate. However, the choice of these four tables is not well justified in the paper.

I agree that this could have been explained better in the paper. In particular, a previous paper (Miller etal, 2007) proposed 12 tables and two of the four tables assessed were not included in the list. Unfortunately, the recoding of some of the variables (for example, qualifications) were not readily available to us and so similar tables were used instead. There are two new tables being assessed for the next paper.

The link between this paper and the first paper on criteria to be used for assessment seems to be a bit weak. The differencing problem doesn't rate a mention in this latest paper except in 'suggestions for further work'.

The next paper will include more developed links between the criteria and analysis. The new tables are specifically chosen to address differencing, and the paper will also include more detailed assessments of additivity and consistency relating to the different methods.

The ORCD and CPCD datasets are different so it's an unfair comparison.

Despite the acknowledged differences between the CPCD and ORCD files, we are assessing the broad statistical effects of the methods as well as the general implications for disclosure risk, rather than comparing like for like. Moreover, the datasets are sufficiently similar to provide a reasonable comparison and will meet the requirements. We need to balance the tremendous amount of resource that would be needed to produce directly comparable datasets and carry out further analysis on them against the gains that would be achieved, and we consider that providing directly comparable datasets would certainly delay the decision on SDC strategy for 2011.

There is a concern that ONS haven't made much progress on the ABS method so cannot compare to this.

At the time of the interim report, we were still gathering information on the ABS method but wanted to get feedback on results so far. We have made good progress here. We are simulating the ABS method and introducing an invariant cell perturbation to preserve additivity while maintaining most of the consistency between tables. We are referring to this as the IACP – Invariant ABS Cell Perturbation method; early results are promising but we will say more about this in the paper due September.

There are cases where record swapping would offer no protection and over-imputation would have not been considered.
There are problems with record swapping such as getting good matched records to swap, especially records which are fairly unique. What about communal establishments (the issue of communal establishments was raised by more than one respondent)?

There will be unusual records that will be difficult to protect (e.g. if there was one person living in a caravan in a small district) whatever method we choose – we would need to take a view as to which standard outputs to produce. We will be looking at communal establishments at a later stage.

On the overall recommendation to drop over-imputation, there was general agreement that it appeared the weakest of the methods evaluated at this stage, but that further evidence needed to be gathered.

Taking into account the arguments in the report, would be hard to argue that over-imputation was preferable with further empirical work. However there are counter-arguments which haven't been proven in the analysis.
I have concerns about dropping over-imputation at this stage.
I have concerns about dropping without comparison to the ABS method.
Further work could be done to improve over-imputation.

It is accepted that we have to carry out further analysis on new tables in one new area. This may provide us with additional evidence to demonstrate that over-imputation performs less well than other methods. Intuitively, over-imputation, particularly targeted, is likely to reduce the diversity of a population, removing 'unusual' observations and replacing them with values from donors more likely to be at the centre of a distribution. We will be comparing the results emanating from over-imputation, and record swapping, to those from the method developed from the ABS in the final evaluation.

Comments suggested that comparisons between targeted and random methods, whether for record swapping or over-imputation, did not provide convincing evidence of any significant difference.

Differences between targeted and random swapping are too small to make a judgement at present.
The reduction in risk from targeting seems negligible to random. Unconvinced by this, although there are a priori reasons that targeting biases estimates.

Though we have considered both targeted and random approaches for the two methods mentioned, at this stage we are more interested in assessing the methods more broadly. If one or both of these methods were selected, making a choice between a targeted and random approach, and the percentage level, would be left for later.

One recommendation in the interim evaluation was that origin-destination tables could not be protected by the SDC methods being considered, at least not without undue damage to data quality. Members of the sub-group agreed.

I think this is a natural recommendation.
I welcome the recognition that SDC will not work for the O/D tables. Good idea to consider licence or access arrangements.

How the O/D tables are protected should be discussed in a different forum as this is much more of a specialised interest with more limited market.
Table 4 is treated as a special case and I agree that O/D tables are deserving of special treatment.
It is important to consider these types of tables as they are the most requested special tabulations; also ONS should consider developing a user-defined flexible table generating package.

A number of users have called for the approach of licensing/access restrictions to address the origin-destination tables and we believe it to be the most reasonable for researchers, while still protecting the data, through access conditions rather than significant data perturbation. ONS Census Outputs are indeed investigating the feasibility of users generating flexible tables from hypercubes, in the light of feedback from a user consultation on outputs. The majority of respondents reported that they would find the facility to generate their own user defined tables very useful and two-thirds said they would prefer to have this facility together with a smaller number of standard pre-defined tables than to have only pre-defined tables equivalent in scope and quantity to 2001 Census tables. Note that there is an EU requirement for ONS to produce some census outputs in the form of hypercubes, and ONS is currently looking at systems to do this.

There were some comments on the analysis of the combination of record swapping with small cell adjustment (SCA), the SDC methods employed in 2001.

The combination of SCA with record swapping is often worse than record swapping alone in the tables of results.
Record swapping comes out as preferable but may not give enough protection to small cells. Thus it would be feasible to consider swapping with SCA.
The aggregated effect of SCA when building from lower level geographies is a disadvantage.

Small cell adjustment does protect the small cells but causes great risk of disclosure by differencing, from experience of 2001. The data utility is also greatly affected by small cell adjustment, especially in constructing bespoke geographies from small geographies where there may be lots of adjusted counts. We are not considering small cell adjustment as a possible method for use in 2011 due to the above points and the unpopularity with users subsequent to the 2001 Census, the reasons for which have been well rehearsed elsewhere. We are running tables through the small cell adjustment method to provide a baseline comparison for other methods.