Progress towards a table builder with in-built disclosure control for 2021 Census

Keith Spicer and Iain Dove, Office for National Statistics (UK).

1. Introduction

This paper reports on progress toward the approach to protect the confidentiality of individual respondents to the 2021 UK Census. This protection is enshrined in law and it is thus work that requires regular updates to UK Census Committee and the National Statistician, who are required to give specific approval. The context of this work involves a reasonable amount of history, lessons learnt from previous censuses, and the satisfaction of an active and vocal user community. The scope of disclosure control for the UK census is wide, but the main focus of this paper is on the protection of confidentiality within tabular outputs. We discuss the methods used in past censuses and introduce a possible approach for 2021 Census.

2. Background

Statistical disclosure control covers a range of methods to protect individuals, households, businesses and their attributes (characteristics) from identification in published tables (and microdata). There is a large literature base now established on disclosure risk, disclosure control and its methodology, notably Hundepool et al (2012). Box 1 highlights the most common forms of disclosure with tabular outputs.

ONS has legal obligations under the Statistics and Registration Service Act (SRSA, 2007) Section 39, and the Data Protection Act (1998) in this respect, and ONS must also conform to the UK Statistics Authority Code of Practice for Official Statistics (2009) that requires ONS not to reveal the identity or private information about an individual or organisation. The Data Protection Act is effectively superseded by the General Data Protection Regulation (GDPR) that comes into force in UK on 25 May 2018. More generally, we have a pledge to respondents on the first page of the census form that the information will only be used for statistical purposes, so we must look after and protect the information that is provided to us. If we do not honour our pledge there is a potential risk that response rates to all our surveys could be adversely affected as could data quality. Moreover, a breach of disclosure could lead to criminal proceedings against an individual who has released or authorised release of personal information, as defined under Section 39 of the SRSA.

The SRSA defines “personal information” as information that identifies a particular person if the identity of that person—

(a) is specified in the information,

(b) can be deduced from the information, or

(c) can be deduced from the information taken together with any other published information.

There are exemptions from the SRSA, through which information can be disclosed, for example where it has already lawfully been made publicly available, is made with consent of the person, or is given only to an approved researcher under licence. Note that it is not a breach under the SRSA to release information that could lead to an identification of an individual, where *private* knowledge is also necessary in order to make that identification.

Box 1. Types of Disclosure

Identification Disclosure: The ability to recognise or identify oneself (or another respondent) as the 1 individual in a table cell. [See Table 1 and the two cells in Very Bad Health column]

Attribute Disclosure (AD): The ability to learn something new about a respondent (or group of respondents) from a table. This is usually where a row or column only has one non-zero entry. [See Table 1 – All Black males have Fair Health]

Within Group Disclosure: A combination of both Identification and Attribute Disclosure. It is the ability to learn something new about a number of other respondents, where a row or column has contains a 1, and only one other non-zero entry. The respondent represented by the 1 can deduce information about the other group members. [Table 1 – the Asian male with Good Health knows all others have Bad Health]

Table 1. Exemplar disclosure table: Ethnic Group x Health (Males)

	Good Health	**Fair Health**	**Bad Health**	**Very bad Health**	**Total**
White	6	7	3	2	**18**
Mixed	2	2	3	1	**8**
Asian	1	0	5	0	**6**
Black	0	5	0	0	**5**
Other	0	0	0	1	**1**
Total	**9**	**14**	**11**	**4**	**38**

In order to remain within the law, the data provider must take account of all reasonable sources that might be used to try and identify an individual. The UK Statistics Authority Code of Practice for Official Statistics (2009) underlines the need for arrangements for confidentiality protection that protect the privacy of individual information but that are not so restrictive as to limit unduly the practical utility of official statistics.

The importance of this work is underlined by the potential sanction within the SRSA: An individual who contravenes the legislation and is convicted, could receive a custodial sentence for up to two years, or a fine, or both. This is a sanction for an individual but a breach would also result in significant reputational damage for ONS, as well as considerable scrutiny from select committees, privacy lobbyists and pressure groups, and the media.

3. Context – previous censuses

The 1920 Census Act was the first legislation to mention the confidentiality of respondents in UK censuses. However, the understanding of the intricacies of statistical disclosure (as opposed to the security of the forms and their information) did not result in any specific disclosure control measures until the 1971 Census. Previously, there had been some protection in tables due to many being based on a 10 per cent sample of respondents. The 1991 Census used a method of cell perturbation referred to as Barnardisation, whereby some cells in some small area tables had random noise added or subtracted.

In the 2001 Census, the records on the output database were slightly modified by random record swapping. This means that a sample of households was 'swapped' with similar household records in other geographical areas. The proportion of records swapped was the same in all areas. No account was taken of the protection provided through differential data quality (due to, e.g. different levels of non-response imputation). Information about the proportion of records swapped cannot be provided as this might compromise confidentiality protection.

Random record-swapping had some limitations and the Office for National Statistics (ONS) became increasingly concerned about these. It was felt that it would not be apparent to a person using the census data that any method of disclosure protection had been implemented. There would be a perception that persons and households were identifiable (particularly for a single count) and the observer might act upon the information as if it were true.

At a late stage (in fact, after all the disclosure control methodology had been agreed and communicated to users) a review was held to decide on the implementation of additional disclosure protection. The decision was to add a post-tabular small cell adjustment (SCA) method. It involved adjusting the values of small cells up or down according to rules that a proportion of the cells with that small value will be adjusted up, while the rest of the cells with that value will be adjusted down. SCA was applied after random record swapping had been carried out on the microdata.
During the process of small cell adjustment:

- a small count appearing in a table cell was adjusted (information on what constitutes a small cell count could not be provided as this may have compromised confidentiality protection)
- totals and sub totals in tables were each calculated as the sum of the adjusted counts so that all tables were internally additive (within tables, totals and sub totals are the sum of the adjusted constituent counts)
- tables were independently adjusted (this means that counts of the same population in two different tables were not necessarily the same)
- tables for higher geographical levels were independently adjusted, and, therefore, were not necessarily the sum of the lower component geographical units
- output was produced from one database, adjusted for estimated undercount, and the tables from this one database provided a consistent picture of this one population.

The fallout from this was considerable. The Office received numerous complaints from users, broadly covering the following:

- The very late decision to implement SCA
- The data looked 'wrong' – in that there were no 1s or 2s and published tables were not consistent with each other
- Consultation with users on this had been limited
- Tables still took time to pass through disclosure checks, since there was a risk of disclosure by differencing
- The method was not harmonised across UK. SCA was employed for tables using data from England, Wales and Northern Ireland while not for Scotland (who felt that the risk was very low anyway).

In 2005, the registrars general agreed that small counts (0s, 1s, and 2s) could be included in publicly disseminated census tables for 2011 Census provided that
a) there was sufficient uncertainty as to whether the small cell is a true value had been systematically created; and
b) creating that uncertainty did not significantly damage the data.

By implication, the uncertainty around counts of 0 in particular corresponds to uncertainty of attribute disclosures.

Additivity and consistency were the key drivers. After a lengthy evaluation, record swapping was chosen as the primary method but targeted to risky records – those records likely to contribute to small cells and attribute disclosures in census tables.

4. Record Swapping – How it Works

Record swapping is now a well established method of disclosure control in scenarios where large numbers of tables are produced from a single microdata source. The US Census employed this for 1990 and all later censuses (see Zayatz, 2003) and its strengths and weaknesses outlined in Shlomo et al (2010), prior to the 2011 UK Census. It has been used in non-census collections (see Kim, 2016) but in the UK its use has predominantly been in the last two national censuses. It is occasionally used on a small purposive scale to protect microdata where there are a small number of very unusual records that require protection. The following describes the method's use within the 2011 UK Census.

Every individual and household was assessed for uniqueness or rarity on the basis of a small number of characteristics (at three levels of geography) and every household given a household risk score. A sample of households was selected for swapping. The chance of being selected in the sample was based largely on the household risk score, so that households with unique or rare characteristics were much more likely to be sampled. However every household had a chance of being swapped. Once selected, another 'similar' household was found from another area as a 'swap'.

The household and its swap were matched on some basic characteristics in order to preserve data quality. These characteristics include household size, so that the numbers of persons and numbers of households in each area are preserved. Households were only swapped within local authorities (LAs) or, in the case of households with very unusual characteristics, with matches in nearby authorities. So there were no households, say, in Cornwall swapped with any in Birmingham.

The precise level of swapping is not disclosed to the public so as not to compromise the level of protection that swapping provides. The level of swapping was lower in areas where non-response and imputation are higher and already provide a degree of protection against disclosure, so the swapping level varied across the UK.

If the level of imputation in an area was high, the level of swapping required was lower than in other areas. We still have to protect the very unusual and more identifiable persons who have completed and returned their census forms, even in the areas with lots of imputed records, so some record swapping was carried out in every area. A consideration for 2021 is that imputation is likely to be improved due to auxiliary information from other sources and so might not provide so much protection.

The swapping methodology is such that every household and every person does have a chance of being swapped, so all cell counts have a level of uncertainty. Indeed, given that some persons do not respond to the census and some questions are not answered by all, there are also imputed records appearing in the census database and therefore in the cell counts. The combination of imputation and swapping produced some apparent attribute disclosures that are not real, and some cell counts that included imputed and/or swapped records.

People or households with rare or unique characteristics might reasonably expect to be able to see themselves or their household in the data. However, there may be a number of reasons why such a person or their household may not be apparent. There is a very small chance that the information may not have been captured properly (especially in paper responses), but more likely the household was selected for swapping with a household in another area, or that it may have been matched with a different household selected for swapping.

No persons or data items are removed from the census data and therefore outputs at national level and high geographies are unaffected by record swapping. The level of non-response and imputation will actually have a far greater effect on any counts seen in the tables than record swapping. Care was taken to achieve a balance between disclosure risk and data utility and, because we are targeting records where the risk of disclosure is greatest, most analyses based on larger numbers was not greatly affected.

Note that record swapping was also applied to communal establishment data. 2011 was the first UK Census in which these were subject to pre-tabular disclosure control. The Frend et al (2011) method was somewhat similar to that for households, where individuals were swapped between communal establishments, with individuals matched on basic demographic characteristics.

5. Assessment of 2011 Outputs post-record swapping

5.1 Assessing Risk in Outputs

The key issue with assessing disclosure risk was that there was no clearly defined measure of what "sufficient uncertainty" was. The agreement of how to measure uncertainty and what level was to be deemed sufficient was only agreed at an extremely late stage. Meanwhile, the output table definitions and layouts were already in development. Agreement with the National Statistician on the criteria to be used was only achieved at a late stage, these being the minimum proportions

of *real* attribute disclosure (AD) cases that imputation and swapping have protected, and

of *apparent* AD cases (i.e. in the swapped data) that are not real.

An intruder testing exercise (see Spicer et al (2013) provided empirical assessments and evidence of the level of disclosure risk, a level that was deemed acceptable in satisfying the need for "sufficient uncertainty".

The result of this was that every table had to be checked against these criteria. The scale of this requirement was enormous, with around 8 billion cells of data released. The number of tables released for 2011 Census was:

229 Detailed Characteristics tables, for MSOA and above (for some it was district and above)

204 Local Characteristics tables, for OA and above

27 Key Statistics tables

75 Quick Statistics tables (univariate), for OA and above

122 various other tables for workday population, workplace population, migrants and others

This total does not include a vast range of origin-destination tables and around 700 commissioned tables to date, the latter still requiring an ongoing SDC resource.

5.2 User Feedback from 2011

- They liked targeted record swapping
- They felt output checking was a bottleneck
- They thought there were "indirect and unintended" consequences of SDC
- Tables were sometimes revised in a way that was not user-friendly
- We were perhaps over-cautious in some situations (e.g. with benign tables age x sex at the lowest geographies)

SDC processing – record swapping - generally went well and we need to build on good practice from 2011. Tables that failed the criteria in 5.1 were re-designed by collapsing categories or raising the geographic level. Re-design caused a delay in the production of detailed tables and frustration among some users about how collapsing had been carried out. It is vital that there are early decisions as to the outputs that ONS is prepared to allow, and the user-defined system should help as a catalyst for that.

Box 2. Why is record swapping not enough? Why can't we just release everything?

The basis of the level of doubt is that a sufficient proportion of real attribute disclosures are removed by imputation or swapping, and a sufficient number of apparent attribute disclosures that are introduced by imputation or swapping. The targeting means the most risky records are much more likely to be swapped. Every household has a non-zero probability of being selected for swapping. Therefore, there is a level of doubt as to whether the value of one is real. It may be that a person has been imputed or swapped so as to appear in that cell, or indeed there may have been another person or persons swapped out so as to move from that cell, thus creating the value of one. So one cannot ever be sure that a value of one that they see in a table is really the true value.

However, in particular cases where tables (or parts of tables) are sparse, it is difficult to protect all the vulnerable cells with an acceptable rate of record swapping (see Table 2). The level of swapping must be kept low enough to avoid significant loss of utility, but it would need a much higher swap rate than would be desirable in order to sufficiently protect the very high numbers of small cells and attribute disclosures. We also have a duty to protect against the perception of disclosure, the perception that we are not properly protecting the data supplied to us by individual respondents. The trade off in maintaining the utility of outputs is therefore to restrict the breakdowns of variables and/or the numbers of cells.

Table 2. Exemplar sparse table: Tenure x Ethnic Group

	White	Mixed	Black	Asian	Other	Total
Owned outright	22	1	0	1	0	24
Owned with mortgage or loan	34	3	0	1	0	38
Shared ownership	1	0	0	0	0	1
Social rented from council	19	0	1	0	1	21
Other social rented	6	0	0	0	0	6
Private landlord	16	0	0	3	0	19
Employer of a household member	0	0	1	0	0	1
Relative or friend of household member	1	0	0	0	0	1
Other	0	0	0	0	0	0
Live rent free	1	0	1	0	0	2
Total	100	4	3	5	1	113

6. The 2021 Census

6.1 Areas for Improvement for Outputs

In its phase 3 assessment of the 2011 Census, the UK Statistics Authority spoke to a range of users about their experience of 2011 outputs. Generally users were positive about the releases and the engagement activities which had been carried out. However concerns were raised around three aspects of dissemination – accessibility, flexibility and timeliness. These findings were consistent with evaluation work carried out by ONS and the other UK Census offices.

The UK Census Offices are determined to build on what worked in 2011 and address what worked less well.

To help focus priorities, early work has looked at a strategy which targets user concern in the three areas highlighted by UK Statistics Authority:

a. Accessibility – Users reported difficulty in locating specific items, in part compounded by the dissemination approach of publishing a high number of predefined tables.

b. Flexibility – Users reported a desire to create their own outputs and frustration with decisions taken on the level of detail made available.

c. Timeliness – Users expressed disappointment that no substantial improvement had been made in 2011 compared to the release of 2001 Census outputs.

In looking again at the process of producing outputs, work is being carried out to evaluate the most appropriate combination of pre and post-tabular methods for disclosure control. The current favoured method is to consider a combination of targeted record swapping along with a post-tabular cell key method.

6.2 The ABS 'Cell Key' Method

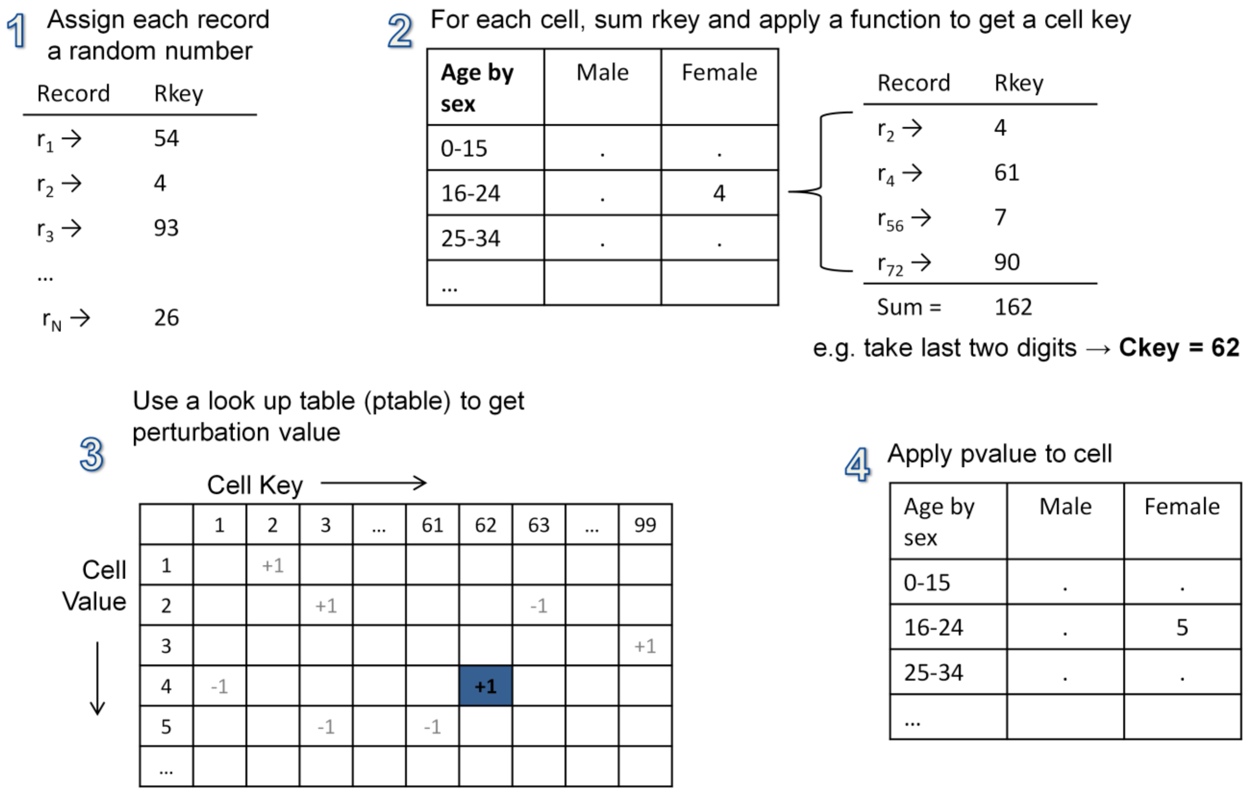
A key part of the work involved assessing the 'cell key' method developed and used at the Australian Bureau of Statistics (ABS). The method is based on an algorithm which applies a pre-defined level of perturbation to cells in each table. The same perturbation is applied to every instance of that cell. In a similar way to record swapping, the precise level of perturbation would need to be set as part of the development of methods (Fraser and Wooton, 2005).

In the lead up to 2011 UK Census, ONS had considered a variant of the ABS method (Shlomo and Young, 2008) but had ultimately rejected it on the basis that it would give rise to small amounts of inconsistency between cell counts and their breakdowns. The inconsistencies would have been small but users had previously expressed their strong desire for additivity and consistency as the most important criteria for 2011 outputs.

The simplest version of the method is demonstrated in Box 3. Every record within the microdata is assigned a record key, which is a random number across a prescribed range, typically 0-99. The random numbers are uniformly distributed. When frequency tables are constructed, each cell has a number of respondents, and the cell key is calculated by summing their record keys. The combination of cell value and cell key is then read from a previously constructed look-up table (termed the ptable) to decide the amount of perturbation that should be used.

Where the same cell (or same combination of respondents) appears in different tables, the perturbation will be the same, due to the same cell value and cell key.

Box 3. Example of the Cell Key Method



1 Assign each record a random number

Record	Rkey
r_1 →	54
r_2 →	4
r_3 →	93
…	
r_N →	26

2 For each cell, sum rkey and apply a function to get a cell key

Age by sex	Male	Female
0-15	.	.
16-24	.	4
25-34	.	.
…		

Record	Rkey
r_2 →	4
r_4 →	61
r_{56} →	7
r_{72} →	90
Sum =	162

e.g. take last two digits → **Ckey = 62**

3 Use a look up table (ptable) to get perturbation value

Cell Key →

Cell Value ↓

	1	2	3	…	61	62	63	…	99
1		+1							
2			+1				-1		
3									+1
4	-1					**+1**			
5			-1		-1				
…									

4 Apply pvalue to cell

Age by sex	Male	Female
0-15	.	.
16-24	.	5
25-34	.	.
…		

The main advantages of the method are that it allows tables to be protected without the need for a case-by-case assessment of disclosure risk and that a greater combination of outputs can be produced. This has potential for a step change in the flexibility of outputs. As demonstrated by ABS, the method can be used to systematically protect user defined outputs. The main disadvantage is that although the same cell of data is consistent in all outputs, there may be differences between that cell and the equivalent aggregation of other cells. Hence the number of 20-24 year olds in Southampton will always be the same across different tables but this may not be the same as the sum of 20, 21, 22, 23 and 24 year olds in Southampton.

There can be an additional protection within the method whereby all 1s and 2s are perturbed, either to 0s or cells of at least size 3. This is not a direction in which ONS should be going, since it resonates of the small cell adjustment method in 2001 UK Census, a method that was deeply unpopular with users. The intention for ONS would be to maintain the appearance of 1s and 2s in output tables, even though many will have been perturbed. It is to be noted that the intended method for ONS SDC is for a light touch cell key perturbation to support the primary method of record swapping.

The light touch of the cell key method should mean that the inconsistencies between different tables are kept to a minimum. It should also mean that most outputs should be available extremely quickly, and not subject to manual case by case checking, as had been the case in 2011. Though there will be differences (inconsistencies) between cell counts and the counts of breakdowns of these cells, the cell perturbation should offer considerable protection against disclosure by differencing. Indeed, when the ABS method was originally proposed, it was principally as a method for protecting against differencing (Fraser and Wooton, 2005).

Users should be able to achieve much greater timeliness by having access to a table builder, within which they can define the tables desired. There is some work for ONS SDC to assess the level of detail available, and this is a function of the parameters relating to the other elements that provide or affect disclosure protection: record swapping, the level (swap rate), the other targeting and matching parameters, and the level of perturbation in the cell key method. This work will enable us to assess how much of the demand for census information can be met using a user-defined table builder system, and how much will need to be serviced through a commissioned tables team.

6.3 Perturbing Zeros

ONS SDC is aiming to apply cell perturbation as a protection against differencing, which is not automatically provided by record swapping. Since differencing is a higher risk for lower geography tables, and unperturbed counts at higher geographies are desirable to users, one could consider an option of leaving higher geography tables without perturbation. The issue with this is it allows comparison of some perturbed and unperturbed values. If for example Local Authority level tables were unchanged, an LA table could be produced (with no perturbation) then compared with the sum of the MSOA counts (and some perturbations) within that LA. In most cases it is not possible to unpick the perturbation and determine the level of perturbation but the exceptions to this are low counts, especially of 1, at the lowest level of geography at which perturbation is not carried out.

This method does introduce uncertainty when attempting to make comparisons between unperturbed counts at one geography and perturbed counts at a lower geography. However, the counts that are both low and known to be unperturbed are the issue, even if the geography is high. An option to counter this, and add uncertainty into any claims of disclosure – notwithstanding the record swapping that has taken place previously – is to allow perturbation of cell counts of zero.

In order to perturb cells with counts of zero there are several differences from perturbing populated cells that need to be dealt with:

- i) For the standard perturbation, the value of perturbation is determined by the cell value, and the 'cell key' which is generated using the record keys of all individuals within the cell. The zero cells contain no records with which to do this.
- ii) Other cell values receive noise that is both positive and negative, ensuring it has an expected value of zero, but since negative counts are naturally not allowed, any perturbation of a zero must be positive, to a one or two, say. This would introduce an upwards bias to the table population by only increasing the cell counts.
- iii) For sparser tables at lower geographies especially, the zero cells make up the vast majority of counts. This means that the frequency table will be sensitive to even low rates of perturbation.
- iv) Some of the cells will be structural zeros, cells which represent a combination of characteristics that are considered highly unlikely to occur, if not impossible. These cells must be kept as zero to avoid inconsistencies, confusion, and user perception of low quality data.

The first issue can be overcome by distinguishing between the zero cells using the characteristics of the cell itself rather than the records belonging to it. We assign a random number to each category of each variable and use the modulo sum of these random numbers to produce a random and uniformly distributed category cell key, in a very similar way to the cell key. This category cell key can be used to make a random selection of cells to perturb. Applying a category cell key in this way ensures zero cells are perturbed more consistently across tables the same way the cell key method ensures consistency when the same cell appears in different tables. This repeatability is obviously preferable to simply selecting random zeros within a table to be perturbed.

The ptable is unbiased in that, for each non-zero count, equal numbers of cell counts are perturbed up as down. In order to provide the protection of perturbing some zeros, we also need to deliberately perturb some additional counts down to zero, and so preserve this unbiasedness. To decide how many additional cell counts of one are perturbed down to zero, there is an algorithm that looks at the numbers of cell counts of zero and one, both at this and higher geographies, to consider the level of disclosure risk present before this extra perturbation. Then the requisite numbers of cell counts of one are perturbed down to zero and an equal number of zero cells are perturbed up to one, using the category cell keys.

Structural zeros (see next section), which should not be perturbed, are given an arbitrarily low category cell key (say 0.001). The cell counts are perturbed to one for the desired number of zero cells with the highest category cell keys. This avoids any population in a cell that has a structural zero count.

6.4 Determining structural zeros

Although structural zeros are well defined by the edit constraints, implementing all constraints in the code would be lengthy, slow to run, and leave margin for human error. Cells that defied any constraints are checked for in all tables (whether or not the edit is relevant) and conditions defined on several variable breakdowns, and potentially millions of possible combinations of categories in different variables. A suggested alternative is to use the cell counts from elsewhere in the table to signal whether the combination should be considered as highly unlikely or impossible. This method allows or disallows the perturbation of a zero cell based on whether that combination has occurred in a different geographical area. This method creates the frequency table at a higher geography (perhaps regional or national level) and assigns a low category cell key to all cells that are zero at that higher geography, i.e. have not been observed elsewhere in the country.

So if a combination of characteristics has occurred elsewhere in the table, it is allowed to reoccur in another area. If a combination has not been observed elsewhere this is prevented from occurring as a result of perturbation. This will cover all cells mentioned by the edit constraints (since they will have been edited out of the microdata before this stage) and other combinations that were feasible but were not observed in any geography (very unlikely to occur).

The main difference caused by this change to the method is that it prevents cases that have never occurred by chance from being perturbed, even if they were not explicitly ruled out by edit constraints. Conversely, it would allow perturbation to occur in a cell that defied the edit constraints, if this combination had occurred anywhere in the data, though since the edit constraints will have been applied to the microdata before this stage, this should not be possible. The method thus allows combinations to happen as long as they remain ‘possible if unlikely’.

Note that this change does not affect the rate of perturbation or how many zeros are perturbed, only the selection of which zero cells are disallowed/excluded from perturbation. In many cases this change has little impact as a zero cell is initially unlikely to be perturbed, equally the cells that would be chosen for perturbation are unlikely to be structural zeros to begin with. An example of the use of the algorithm for perturbing zeros is outlined in Box 4.

Box 4. Exemplar use of the perturbing zeros algorithm (Table = Age x Marital Status)

Step 1. Assign category keys to variables.

Age	Category key
0-15	0.924
16-24	0.864
25-34	0.336
…	
Marital Status	Category key
Single	0.484
Married	0.732
Divorced	0.111

Step 2. For each ‘zero’ cell, calculate Category Cell Key = sum of category keys for that cell

Age by Marital Status	Single	Married	Divorced	Category	Category Cell Key
0-15	14	0	**0**	Age: 0-15	0.924
16-24	8	4	0	Marital Status: Divorced	0.111
…	…	…	…	Sum Category-key = 1.035	
				Cell key = 1.035 mod 1 = 0.035	

Step 3. Where cell count is 0 even at higher geography, assume ‘structural zero’ and assign Category Cell Key as insignificant low value (in this case of ‘0-15 Divorced’ replace 0.035 by 0.001).

Age	Marital Status	Category Cell Key	Cell value	Higher geog cell value
0-15	Single	.	14	223
0-15	Married	0.001	0	0
0-15	Divorced	0.001	0	0
16-24	Single	.	8	151
16-24	Married	.	4	77
16-24	Divorced	0.975	0	2
…	…			

Step 4. Calculate how many zero cells need to be perturbed. Perturb those with the higher or highest Category Cell Keys.

Age	Marital Status	Category Cell Key	Cell value
0-15	Single	.	14
0-15	Married	0.001	0
0-15	Divorced	0.001	0
16-24	Single	.	8
16-24	Married	.	4
16-24	Divorced	0.975	0 → 1
…	…		

7. Summary

This paper has set out the need for disclosure control, putting into context alongside the history of data protection in previous censuses. The desire is to move to a user-defined table builder system that services a high percentage of user demand, and work continues to develop the methodology to do that. Alongside targeted record swapping, that was used in 2011 and is being enhanced, the basis of the post-tabular methodology has been previously developed by ABS, with enhancements to make this suitable for ONS to use in the 2021 Census. In particular, the provision of low counts must be supported by an algorithm to perturb zeros, which protects against 'disclosure by existence'. The weaknesses of the previous census, including flexibility and timeliness, are addressed head on by such a system, though balanced by a small amount of inconsistencies between different tables.

Whilst further developing our thinking and methodology, we are engaging further with users to assess their appetite for such a facility, and to maximise the amount of information that can be gleaned from such a table builder. Work is currently taking place on the business rules that are required to decide which combinations of variables, categories and geographies will be permitted.

References

Andersson, K., Jansson, I. and Kraft, K. (2015) Protection of frequency tables – current work at Statistics Sweden. *Joint UNECE/Eurostat work session on statistical data confidentiality* (Helsinki, Finland, 5-7 October 2015).

Fraser, B. and Wooton, J. (2005) A proposed method for confidentialising tabular output to protect against differencing. *Joint UNECE/Eurostat work session on statistical data confidentiality* (Geneva, Switzerland, 9-11 November 2005).

Frend, J., Abrahams, C., Groom, P., Spicer, K., Tudor, C. and Forbes, A. (2011) Statistical Disclosure Control for Communal Establishments in the UK 2011 Census. *Joint UNECE/Eurostat work session on statistical data confidentiality* (Tarragona, Spain, 26-28 October 2011).

Hundepool, A. Domingo-Ferrer, J., Franconi, L. Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.P. (2012) *Statistical Disclosure Control,* Wiley Series in Survey Methodology.

Kim, N. (2016) The Effect of Data Swapping on Analyses of American Community Survey Data. *Journal of Privacy and Confidentiality* 7(1); 1-19.

Shlomo, N., Tudor, C. and Groom, P. (2010) Data Swapping for Protecting Census Tables. In *PSD 2010 Privacy in Statistical Databases*. Germany: Springer LNCS 6344; p41-51.

Shlomo, N. and Young, C. (2008) Invariant Post-tabular Protection of Census Frequency Counts. In *PSD 2008 Privacy in Statistical Databases*. Germany: Springer LNCS 5261; p77-89.

Spicer, K., Tudor, C. and Cornish, G. (2013) Intruder Testing: Demonstrating practical evidence of disclosure protection in 2011 UK Census. *Joint UNECE/Eurostat work session on statistical data confidentiality* (Ottawa, Canada, 28-30 October 2013).

Zayatz, L. (2003) Disclosure Limitation for Census 2000 Tabular Data. *Joint ECE/Eurostat work session on statistical data confidentiality* (Luxembourg, 7-9 April 2003).

Relevant Legislation

Data Protection Act (1998) http://www.legislation.gov.uk/ukpga/1998/29

General Data Protection Regulation (in force in UK May 2018): Overview.
https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/

Statistics and Registration Service Act (2007)
http://www.legislation.gov.uk/ukpga/2007/18/section/39

UK Statistics Authority Code of Practice for Official Statistics (2009)
https://www.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-codeofpracticeforofficialstatisticsjanuary2009_tcm97-25306.pdf