**Statistical disclosure control (SDC) for 2021 UK Census**

**Keith Spicer, ONS SDC Methodology Branch.**

1. **Introduction**

This paper reports on progress toward the approach to protect the confidentiality of individual respondents to the 2021 UK Census. The paper builds on previous papers and is an evolving working document as progress is made. 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 balanced against the legal requirements of Office for National Statistics (ONS) to protect the confidentiality of individual respondents. 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 preferred approach for 2021 Census.

1. **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 The Data Protection Act is effectively superseded by the General Data Protection Regulation (GDPR) that came into force in UK on 25 May 2018.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 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.

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

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]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **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 of 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.

1. **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.

 **3.1 2001 Census**

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). Further 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 manual table checking, 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.

**3.2 Evaluation pre-2011**

The SDC team had previously undertaken a long post-2001 Census evaluation review to assess all possible disclosure control methods that could be used for protecting frequency tables in the (forthcoming) 2011 Census. ONS (2010) describes this evaluation, the first phase of which narrowed down the methods to a short-list of three: record swapping, over-imputation, and an Invariant ABS Cell Perturbation (IACP) method. The IACP method (see Shlomo and Young, 2008) was an extension of the Australian Bureau of Statistics (ABS) approach where controlled noise was added to cell counts in frequency tables.

Users’ uppermost concern was that the tables were additive and consistent. Given these were key issues within the negative feedback from 2001, and that additivity and consistency could not both hold for the IACP, that method was discounted as failing on the mandatory criteria. The remaining two methods, record swapping and over-imputation, were scored against a number of criteria developed by ONS SDC and the UK SDC Working Group, which included membership from both GROS and NISRA. The Group agreed the criteria and relative weights, and scored record swapping as slightly better than over-imputation.

1. **Record Swapping – How it Worked in 2011**

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 to 2010[[1]](#footnote-1) (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 included household size, so that the numbers of persons and numbers of households in each area were 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 was 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 had 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 might be improved due to auxiliary information from other sources and so might not provide so much protection.

The swapping methodology was such that every household and every person did have a chance of being swapped, so all cell counts had a level of uncertainty. Indeed, given that some persons did not respond to the census and some questions were not answered by all, there were 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 were 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 was 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 another household that had been selected for swapping.

No persons or data items were removed from the census data and therefore outputs at national level and high geographies were unaffected by record swapping. The level of non-response and imputation actually had a far greater effect on utility than did 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.

1. **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 “sufficient uncertainty”. The agreement of how to measure uncertainty and the level 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”. However, the exercise did highlight some areas within the data that appeared to be vulnerable and these were addressed within the output specifications.

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 900 commissioned and bespoke tables to date, which still require an ongoing SDC resource.

**5.2 User Feedback from 2011**

Though most users of census data appreciate the need for disclosure control to protect confidentiality, the balance between the needs for protection against disclosure risk and for providing sufficient data utility creates a tension between data providers and end users. In commenting on the user needs in relation to this, the user feedback generally covered the following points:-

* They liked targeted record swapping as a disclosure control method
* 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
* ONS 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 sometimes 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.

1. **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.

The challenge of satisfying the user community is balanced against the legal obligations to protect against disclosure risk. This balance of risk versus utility is the classic problem for statistical disclosure control.

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.

Timeliness and flexibility can be addressed through the availability of an on-line table builder, allowing a user to define the table that they require. The level of detail that a user can be allowed is subject to the assessment of disclosure risk that that combination of variables and geography will generate. This has sometimes been referred to as ‘disclosure control on the fly’ but much of the risk assessment has to be carried out prior to the table builder being available. This is discussed at length later in this paper (Section 7).

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 |

**6.2 Changing landscape**

After the 2011 Census, feedback was that users generally liked the record swapping method, but the key demands for a future census were for flexibility, accessibility and timeliness, above additivity and consistency. At the same time, changes in the data environment, a greater public awareness and concern over data privacy and security, and more computing power for more sophisticated intruders has necessitated consideration of improved methods for assessment and protection of risk.

On the utility side, the need was for a disclosure control method that would allow a table builder with users able to interact by requesting combinations of census variables to form tables. The potential methods were:

* As is 2011 (targeted record swapping)
* More (a higher level of) record swapping
* ABS Cell Key method
* Record swapping plus ‘light touch’ cell key perturbation
* Small cell adjustment

**6.3 Differential privacy**

An emerging approach to protecting individuals within datasets is that of differential privacy. The overriding principle is that “the risk to one’s privacy …. should not substantially increase as a result of participating in a statistical database” (Dwork, 2006). Effectively, one should not be able to learn anything about an individual by differencing between outputs from a database containing the individual and outputs from the same database that does not contain the individual. Put differently, differential privacy hides the presence of an individual in the database from data users by making two output distributions, one with and the other without an individual, be computationally indistinguishable (for all individuals) (Lee and Clifton, 2011).

Of course, the risk here is not zero, but should be very low. The approach is characterised by the parameter Ɛ to quantify the privacy risk posed by releasing the outputs. Ɛ reflects the relative difference between the probabilities of receiving the same outcome from the two different datasets, one with and one without the individual subject. A release that satisfies the criterion of being bounded above by Ɛ is said to be Ɛ-differentially private.

Differential privacy is not a method but an approach. Differential privacy typically uses an output perturbation technique which adds random noise to the outputs. The selection of Ɛ determines the level/distribution of that noise, which may be fairly arbitrary and very much driven by the level of risk aversion of the data provider. As in traditional disclosure control approaches, the goal is to find a good balance between reducing risk and maintaining usefulness of the end use release. The field of differential privacy is still young and a too stringent choice of Ɛ generally eliminates any useful information. However, there are some specific use cases where differential privacy can be used without any significant loss in utility (Cabot, 2018).

The approach is mathematically neat but mostly still at the theoretical and academic stage. The use is not yet developed for general-purpose practical applications or for an NSI yet to use in anger in a national census. US Census Bureau are aiming to use for 2020 US Census and Garfinkel et al (2018) provide a useful discussion on issues encountered towards that aim. With differential privacy still in its infancy, the tendency is to over-protect the data taking a “worst case” scenario, though this is intentional to design against a potentially sophisticated intruder taking advantage of any vulnerabilities that may lie within the outputs. Rinott et al (2018) discusses the use of differential privacy in protecting frequency tables, in particular the issues in assessing the appropriate level and distribution of noise.

Some traditional SDC methods have elements of differential privacy in them, though perhaps not so formal. The level of noise added by perturbative methods is not normally parameterised as formally as in a differential privacy approach. Moreover, bearing in mind the main protection of record swapping in the 2021 UK Census case, there are useful aspects of differential privacy that can be considered. In particular, it would be helpful in assessing the level of parameterisation that would be needed in the cell key method, discussed in section 6.4, which is effectively a flavour of the differential privacy approach.

UK Census Committee (UKCC) advised that development work should concentrate on the option of Record swapping plus ‘light touch’ cell key perturbation.

**6.4 The ‘Cell Key’ Method**

A key part of the work towards 2021 involved assessing the ‘cell key’ method developed and used at the 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 here as 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.

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 an area 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 other tables.

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. However,this would resonate of the small cell adjustment method in 2001 UK Census, a method that was deeply unpopular with users. The method also only partly protected the outputs due to a residual risk of disclosure by differencing. A better 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 in 2021 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).

Disclosure by differencing occurs when two or more separate outputs, each of which are ‘safe’ in isolation, can be used in combination to produce a disclosure. For instance, a table based on ‘all adults aged 16-64’ could be compared with an identical one based on ‘all adults aged 16-65’, such that the subtraction of counts from the former from those of the latter would give a table relating just to the likely small number of those adults aged 65. Differencing is most likely to occur due to either the use of marginally different breakdowns of the same variable or the use of overlapping geographies giving rise to geographic ‘slivers’. Where noise is added through the cell key method, the difference between two counts has some uncertainty, since either, both or neither of the counts may have been perturbed. Therefore, some of the apparent differences may not be real, and it is important that there is sufficient metadata so that a potential intruder is aware of this potential.

Box 3. Example of the Cell Key Method



**6.5 Perturbing Zeros**

ONS SDC is aiming to apply cell perturbation as a protection against differencing, which is not automatically provided by record swapping. Since disclosure by 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 known unperturbed values. If for example Local Authority (LA) 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 many 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. Even then, there is almost always uncertainty as to the perturbation present within the cell counts of the lower geographies.

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 (and therefore no record keys) 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) should not be perturbed, and so 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.

**6.6 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 region/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, e.g. a 90 year old student).

The main difference caused by this change to the method is that it prevents perturbation of zeros for combinations of attributes that have never even occurred anywhere else, even if they were not explicitly ruled out by edit constraints. Conversely, it would allow perturbation to occur in a cell, if this attribute combination had occurred anywhere in the data. The method thus allows attribute 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 |
| … | … |  |  |

1. **Flexible Table Builder**

The intention is to have a facility for users to build their own bespoke tables, selecting the desired combination of variables and choice of breakdowns within them. There will be some restriction on these, and it is a moot point as to how ‘flexible’ one might consider that to be, given that there is a finite (though extremely large) number of tables that can be built. The Table Builder will be dynamic in the sense that tables will be assessed for disclosure risk at the time of request – rather than as one of a previously defined list to be assessed prior to the release of the Table Builder. That said, there will be a number of ‘pre-canned’ combinations of variables that will be available as the user requests popular tables.

**7.1 Risks – recurrent unique combinations**

Having users able to produce their own tabulations creates specific disclosure risks. The most pertinent is where an individual is unique on one or a small number of variables and repeated similar requests could expose the whole of an individual record. For example, if there is only one person in an area with a specific attribute (a unique value for variable A), then requests of A x B, A x C, A x D, etc. will build up the census record relating to one individual – the respondent who is unique in the category within variable A. This risk is addressed in three ways, (1) targeting record swapping more heavily towards all ‘risky records’, (2) adding noise through the cell key method – see 6.4 - and (3) restricting the tables which users can build. We have termed the latter ‘Business Rules’ or ‘Disclosure Checks’.

**7.2 More Targeted Swapping**

Protection for the 2011 Census data concentrated on targeted record swapping. A small number of variables were selected on the basis of being visible, sensitive or likely to be known by friends or associates. Every individual was assessed for disclosure risk by considering whether they were unique on any of those variables at OA, MSOA and LAD level. Any household containing one or more ‘risky’ individuals was deemed a risky household.

Households were selected for swapping from the pool of ‘risky’ households, the proportion swapped in any area based on the prevalence of risky records, the population size, and the level of record imputation. There was also a smaller proportion of ‘non-risky’ households that were swapped. Every swapped record was matched with another record in another area at the record’s lowest non-unique geography (e.g. a record risky at OA but not at MSOA was swapped with another record outside the OA but within the MSOA); these were matched on household profiles that included basic demographic variables. At the very least, households were matched on household size so as to not affect total population size, except in the cases of large households (size 8+) where the sizes may have been slightly different. Some effort was made to match a risky record in one area with another risky record in another area.

Given the specific risk scenario outlined in the previous section, it is proposed that we only swap risky records, and we swap all records that are risky based on a much larger number of variables. We propose to swap all OA-risky records between different MSOAs to increase the uncertainty that a very near neighbouring OA could house a risky individual. Finding households with which to match could lead to non-risky households being swapped with those risky households. As in 2011, this means that every household has a chance of being swapped.

**7.3 Disclosure Checks (Business Rules)**

In making possible the construction of tables from an on-line table builder, it is important to note that this does not mean that every combination of variables will be available for every level of geography. Indeed, if that were the case, we would effectively be providing individual level microdata where the perturbation could be unpicked from the billions of different combinations of variables and variable breakdowns. Albeit that the resultant microdata would be post-swapping, the perception would (correctly) be that we would be providing personal information for every census respondent, though some might not be quite in the right geographic area. Certainly it would be straightforward to identify individuals from knowing a few of their details and roughly where they live, and thus discover the remainder of their information.

The disclosure checks are the rules by which decisions can be made as to whether to allow release of outputs pertaining to specific combinations of variables. In previous censuses, the policy has always been to assess whether the release of tables is acceptable for all areas, and so every table that was passed was available for every area. That did mean that tables that might have been acceptable for some areas were not released because the corresponding table was not acceptable for other areas. This was particularly the case for some tables with ethnic group or country of birth, where such minority populations might be clustered in a small number of areas. Our aim for 2021 is to make tables available for those areas where the disclosure risk would be sufficiently low, rather than reject for all areas because some would incur higher risk. We refer to the two approaches as the ‘*blanket*’ approach – where tables are produced for all areas, and the ‘*patchwork*’ approach – where tables are produced for the subset of areas where the risk is low. Note that we can set parameters that define this risk, and so these disclosure checks can be applied by the online table builder, rather than manually, which is a major benefit and hence outputs can be published much faster this time.

We will still be making many oft-requested tables available on a blanket basis through the Table Builder, if available in previous censuses, due to the greater protection of swapping and the cell key method in 2021. The rationale is that if they were sufficiently protected in 2011, they will be at least as well protected in 2021. The patchwork approach mostly covers combinations of variables and categories that were not published in 2011, and provided subject to one or more rules. The types of rules we are considering are listed below. The parameters have not been set and we need to take into account uncertainty introduced by imputation and swapping. Work is ongoing on considering which rules are appropriate and the parameters within them.

Maximum of [x] variables (excluding geography) in the table

Maximum of [x] internal cells (including geography) in the table

All marginal totals in table must be either at least [x] or 0

At least [x] cells in the marginal total must be populated

The maximum marginal total must be at least [x] records less than the table total

The maximum marginal total must be at most [x]% of the table total

No more than [x]% of internal cells can be equal to zero

No more than [x]% of populated internal cells can be a ‘real’ one

No attribute disclosure value can be fewer than [x]

The sum of 'real' attribute disclosure cells, with a value fewer than 3, should be no more than [x]

No more than [x]% of cells can contain a 'real' attribute disclosure with a value fewer than 3

**7.4 Transferability**

 **7.4.1 Administrative Data**

The use of data from administrative sources will be a feature in the outputs from both the traditional census and in the administrative data census in 2021. Firstly, the traditional census will also be supplemented by a number of variables that will be linked/matched to the census microdata. Where outputs include both those from the main questionnaire and those from other sources, these are referred to as enhanced outputs.

The protection of these enhanced outputs will be within that of the table builder: the targeted record swapping followed by the cell key perturbation. In many cases, the linked source will have partial coverage within the enhanced dataset, and the data quality may be lower than in the census. Considering that, the disclosure risk should increase only negligibly with the additional variables. However, there could be a significant increase in risk if the external source data are wholly or partly in the public domain. If that were the case, it may be possible for an intruder to assign an identifier from the external source to a cell count of 1 that transpires from the table builder, and subsequently find out further information about this identified person by requesting other tables, perhaps one variable at a time. Alternatively, it may be possible at low geographies for an intruder to unpick parts of the record swapping by noting existence of an unusual combination of categories within the variables in the public domain. This may be addressed by applying coarser coding on these variables within the table builder so that it is not possible to identify the individual from it.

Though it is probable that a table builder could be used in some shape or form to satisfy the output needs, the admin data census will require a fresh disclosure risk assessment. The disclosure control package for the traditional census is proposed as targeted record swapping as the primary protection method, supported by a secondary light touch cell key method, but work is required to assess the needs with admin data, which may have differential quality. The admin census may well require different protection methods due to different risk scenarios, and even if the admin census were to use record swapping and the cell key method for the “admin census table builder”, the parameterisation will need a separate assessment. It is likely that the parameters will be quite different to those used for the traditional census, and establishing these will require significant input from the SDC team. However, one consideration is to ensure that the swapping cannot be unpicked by comparison of the enhanced outputs with those from the admin data census, since both will contain the similar ‘admin data’ variables. How much the cell key method can protect against this comparison (or whether the swapping can be carried out on the same records in the two sources) is still to be investigated.

 **7.4.2 Other Datasets**

More generally, there is great potential for the table builder to be employed for other data collections and outputs. However, it must be stressed again that this is not a ‘plug and play’ method of protecting data from any source. Its use for any datasets depends on the parameterisation specific to the data. Aspects that might influence the type and level of protection required include: data quality, sampling fraction, coverage, level of imputation, age of the data, sensitivity and risk appetite.

1. **Harmonisation**

ONS, NRS and NISRA are broadly harmonised on the use of targeted record swapping and the cell key method. However, there will be differences between the countries on some of the specific details within the approach. Known and likely variations are:-

* The variables used to assess risk prior to the swapping procedures are likely to differ slightly (to take account of the different questions asked and population differences), as are the variables and profiles for matching to find a household with which to swap.
* When the record swapping is run, it is likely that new ‘uniques’ will be created – records that are still in their true output area (they may originally have been one of two or three with a specific value in variable A, and the others may have been swapped on the basis of uniqueness in other variables). ONS is considering whether to run a second round of swapping targeting these. Current investigation of the uniques swapping code would indicate that due to the high swapping rate it is unlikely that NRS will do even one round of this, and are still planning to run targeted record swapping similar to that run in 2011.
* The three Offices will each set their own perturbation rate within the cell key method (though they *may* be harmonised or similar). In particular, it is likely that the specific rates for perturbing 1s and 2s will differ.
* ONS and NRS aim to apply the methods within the flexible table builder, while NISRA’s intention is to apply the methods to pre-defined data cubes.
* The geographic level above which all tables would be provided unperturbed is still under investigation. It is a working assumption that this will be equivalent to local authority level but it is possible that different countries will make different decisions here.
* There has been no agreement or decision as yet as to whether to perturb totals separately or to make the totals the sum of the perturbed cells.
* The disclosure checks (business rules) in the table builder will almost certainly differ. This aspect is still very much under investigation. ONS is likely to have some pre-defined queries available for all areas but most queries will be subject to the ‘patchwork’ approach, where the output will only be available for areas that pass the business rules. NRS is likely to use rules based around sparsity, number of variables and limit specific combinations, and be based on the ‘blanket’ approach.

**9. Other Census products**

This paper has so far not covered any products outside the frequency tables for combinations of the standard and derived census variables. However, we are aware that there will be different strategies for a number of other products for 2021 UK Census for which there is a user need.

These include:

* Microdata samples. These are now a well established output from the decennial census, with 1991, 2001 and 2011 samples produced at the time, as well as the samples derived from 1961, 1971 and 1981 Censuses as part of the Enhancing and Enriching Historic Census Microdata (EEHCM) project <https://www.ukdataservice.ac.uk/about-us/our-rd/historic-census-microdata> now held at UK Data Service. There is an established Census Microdata Working Group – with external representation - whose second meeting was in August 2018. The general feeling in Census is to have similar products to those of 2011 and, given the general satisfaction of users, to employ similar disclosure techniques. The level of disclosure risk will vary from negligible (for the public ‘teaching’ dataset) to different balances of risk/utility based on access and licensing conditions.
* Origin-destination tables. These are extremely difficult to protect using only traditional disclosure control methods. In 2001, the small cell adjustment method was employed, with all tables being made freely available, and resulted in very poor utility for most users using low level geographies. In 2011, the tables were separated into those available publicly, safeguarded licence and in a secure setting. The tables thus had high utility but access was often more difficult. The approach for 2021 has to address both of these issues. There is an established Census Origin-Destination Working Group – with external representation – whose second meeting was in August 2018. The current thinking is to use a ‘patchwork’ approach to supply breakdowns for those flows in areas that achieve defined thresholds. SDC is presenting this approach at the Privacy in Statistical Databases conference in Valencia in September 2018 (Dove et al., 2018).
* Tables for ‘difficult geographies’ that are not suitable for the standard ‘best fit’ approach. The ONS Geography Policy is for all outputs to be ‘best fit’ to output area geography. This addresses the risk of creating geographic slivers between bespoke and standard geographies. However, there are exceptions where the best fit approach is less suitable. The two most apparent are National Parks, where there are instances of large differences between populations relating to the ‘exact’ boundaries and the ‘best fit’ geography; and parishes, where many are smaller than output areas (and where the desired volume of outputs is small). The solution to the former is straightforward, as in 2011, where exact fit counts and a small number of breakdowns were supplied for National Parks. The latter will require a new approach, and ONS Geography’s current preference is to group parishes into ‘parish councils’ (some of which are already established locally) that are above a population threshold.
* Workplace/workday tables. One of the successes of 2011 was the advent of workplace zones to help provide a more appropriate geography for workplace based statistics, where ONS SDC worked closely with ONS Geography and University of Southampton to address disclosure risk. In particular this benefited some of the origin-destination outputs. For 2021, it is a working assumption that workplace zones will be updated to reflect rises, falls and movements in the workplace populations in these areas.
* Small populations. These are special sets of tables produced for populations that are not available in standard tables or in the Flexible Table Builder. These are predominantly for individuals from ethnic groups, religions and countries of birth who may be clustered geographically around England and Wales. Examples from 2011 Census were Kashmiri, Ravadashian, Albanian COB, Kuwait COB. Analogous to the patchwork approach, the breakdowns are only released for areas that have population above given thresholds: e.g. MSOA100, MSOA200 and LA100.
* Grid square outputs. The anticipation is that these will follow the Eurostat requirement for a population count and a number of univariate outputs for 1 km squares. SDC will have to provide advice and guidance for sparsely populated areas where a 1 km square may only have a small population. Census and ONS Geography are considering whether to provide information for smaller grids where population is most concentrated.
* Commissioned tables service. The Table Builder will not encompass all the possible tables from the census database. In particular, where the population select is rather esoteric, there will be a significant need for a service to produce bespoke or commissioned tables. It is imagined that the need will not be as great as for previous censuses, though making greater numbers of tables available for more areas may actually stimulate greater interest in census products. The SDC team will need to produce advice and guidance for those bespoke tables.
1. **Summary**

This paper has set out the need for disclosure control, put into context alongside the history of data protection in previous censuses. To help focus priorities on outputs and the disclosure control methodology for 2021 Census, development work has considered a strategy which targets user concern in the three areas of outputs highlighted by UK Statistics Authority after 2011:

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.

Timeliness and flexibility can be addressed through the availability of an on-line table builder, allowing a user to define the table that they require. The level of detail that a user can be allowed is subject to the assessment of disclosure risk that that combination of variables and geography will generate. UK Census Committee (UKCC) has agreed that the favoured option to be researched was a combination of pre-tabular targeted record swapping and a post-tabular cell perturbation (cell key) method. While record swapping is the main protection for outputs in 2021, the cell key method offers additional considerable protection against disclosure by differencing and thus allows the desire for user defined on-line tables to be realised.

Alongside targeted record swapping, that was used in 2011 and is being enhanced, the basis of the post-tabular methodology had been previously developed by Australian Bureau of Statistics (ABS), with enhancements developed by ONS SDC Methodology to make this suitable for use in the 2021 UK Census. In particular, the provision of low counts must be supported by an algorithm to perturb zeros, which protects against ‘disclosure of 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. This approach has been shared with the Eurostat SDC Expert Working Group and adopted as the recommended approach for the 28 EU NSIs for protecting hypercubes from the 2020/2021 Census round (Giessing and Schulte Nordholt, 2017).

Whilst further developing our thinking and methodology, Census Outputs and SDC 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 disclosure checks that are required to decide which combinations of variables, categories and geographies will be permitted. A possible approach that is being assessed for 2021 is to allow each output to be provided for those areas for which the data are sufficiently low risk (a 'patchwork' approach), as opposed to a blanket 'everywhere or nowhere' policy employed in previous censuses (an example here is a table using detailed ethnic group that might be available in some areas in cities with greater ethnic diversity). This should facilitate greater amounts of information to be disseminated where previously blocked.

**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).

Cabot, C. (2018) Differential Privacy Re-examined. Available at: https://www.privitar.com/listing/differential-privacy-re-examined

Dove, I., Ntoumos, C. and Spicer, K. (2018) Protecting Census 2021 Origin-Destination data using a combination of Cell-key Perturbation and Suppression. In *PSD 2018 Privacy in Statistical Databases*. https://unescoprivacychair.urv.cat/psd2018/index.php?m=program

Dwork, C. (2006) Differential Privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) Automata, Languages and Programming. ICALP 2006. *Lecture Notes in Computer Science*, vol 4052. Springer, Berlin, Heidelberg.

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).

Garfinkel, S.L., Abowd, J.M. and Powazek, S. (2018) Issues Encountered Deploying Differential Privacy. In *2018 Workshop on Privacy in the Electronic Society* (WPES’18), October 15, 2018, Toronto, ON, Canada, Jennifer B. Sartor, Theo D’Hondt, and Wolfgang De Meuter (Eds.). ACM,

New York, NY, USA. https://doi.org/10.1145/3267323.3268949

Giessing, S. and Schulte Nordholt, E. (2017)  *Development and testing of the recommendations; identification of best practices.* Harmonised protection of census data in the ESS; Work Package 3, Deliverable D3.3 Available at: <https://ec.europa.eu/eurostat/cros/system/files/recommendations_for_the_protection_of_hypercubes.pdf>

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](http://repository.cmu.edu/jpc/vol7/iss1/3/). *Journal of Privacy and Confidentiality* 7(1); 1-19.

Lee J., Clifton C. (2011) How Much Is Enough? Choosing ε for Differential Privacy. In: Lai X., Zhou J., Li H. (eds) Information Security. ISC 2011. *Lecture Notes in Computer Science*, vol 7001. Springer, Berlin, Heidelberg

ONS (2010) Evaluating a statistical disclosure control (SDC) strategy for 2011 Census outputs. Available at: <https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howwetookthe2011census/howweplannedfordatadelivery/protectingconfidentialitywithstatisticaldisclosurecontrol/sdc-evaluation-for-2011-census-tabular-outputspublicfinal_tcm77-189751.pdf>

Rinott, Y., O’Keefe, C., Shlomo, N. and Skinner, C. (2018) Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science* 33(3); 358-385.

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>

1. Note that US Census Bureau are considering an alternative approach for 2020 Census, consistent with the differential privacy paradigm. This is along the lines of adding noise to cell counts in tables. [↑](#footnote-ref-1)