Statistical Disclosure Control update and Differential Privacy suitability assessment

Executive Summary

This paper is an update on our Disclosure control proposal for Future Population and Migration Statistics (Dove 2024) [1]. Updates are provided in particular on the application of swapping and use of record keys. This paper also discusses our current position on the suitability of differential privacy for future population migration statistics. Cell key perturbation has many properties of differential privacy, but it is not expected that perturbation meets the formal definition of differential privacy, particularly on small cells. Further work will investigate whether our current approach is epsilon-delta differentially private, although it may be difficult to stay within a fixed ‘privacy budget’ if multiple sets of users have access to the data.

We are asking for endorsement of our approach of pursuing record swapping and perturbation, and recommendations of any factors we have not considered.

Introduction

All outputs from the Office for National Statistics (ONS) are subject to the Data Protection Act 2018 and the Statistics and Registration Services Act 2007, and so must not contain information that identifies an individual, household or business. Our responsibility to protect confidentiality is also made clear in the UK Statistics Authority's Code of Practice for Statistics [2].

A number of statistical disclosure control methods and processes are used to protect ONS outputs, including those in tabular form and record-level data. These methods introduce uncertainty in the values or combine categories, and while this provides disclosure protection, it also reduces the utility of the outputs (Hundepool, A. and others (2012), 'Statistical Disclosure Control') [3]. There is a balance to strike between the two, usually based on a qualitative judgment of risk. For example, Census 2021 outputs had several disclosure control methods applied to reduce the risk of identification and disclosure, namely targeted record swapping, cell key perturbation, and disclosure checks.

This paper is an update on our disclosure control proposal for future population and migration statistics. The FPMS programme covers a range of population outputs, and our focus will be on the data referred to as ‘admin based census - in development’, a dataset attempting to capture all residents in England and Wales, and their attributes. The ‘Admin based census in development’ produced in August 2024 contained single year of age, sex, and geography at LSOA level. Which attributes will be included in future iterations is subject to change and data availability, but we will be assuming that data will initially contain a small number of attributes, with more attributes being added over time.

For outputs from data including only age, sex, and geography, cell key perturbation will be applied before data is taken out of the secure environment. For future iterations of data including attributes, we also intend to apply targeted record

swapping to the microdata, as well as the cell key perturbation method to frequency tables produced from this data.

This paper includes findings from a literature review of other NSIs approach to disclosure control, with many using a threshold and perturbation or rounding as their primary methods. ONS intends to enable the use of small cell values by introducing sufficient uncertainty in them via our disclosure methods. Various disclosure methods were examined by a Eurostat working group (Nordholt and others 2024) [4]. A combination of targeted record swapping and cell key perturbation remains the recommended approach by Eurostat. (CROS, [5])

A method of producing record keys from ID numbers is discussed, as well as some findings from our record swapping development, particularly the impact of geography information being at LSOA level, and having fewer attributes with which to measure uniqueness used for targeting.

This paper also includes an assessment of our current position on differential privacy for FPMS. Although we do not deem the methods we are applying to meet the definition of differential privacy, the cell key perturbation in particular has very strong similarities and offers a similar mode of protection: Laplacian noise added to frequency tables before their release. We do not intend to apply a strict epsilon differentially private system, to avoid potential issues with the utility of outputs. Differential privacy is best suited to a fixed set of outputs, known ahead of time. The admin-based census data will be subject to further research projects over a longer time scale, from a variety of users. As such, it would be very difficult to anticipate what information will be produced from the data ahead of time and ensure the set of outputs as a whole is differentially private, and stays within a ‘privacy budget’. Further research and trials of differential privacy will be carried out in future work.

Literature review

One of the next steps in our work was to carry out a review of how other NSIs are approaching disclosure control. Our findings were that the most commonly used methods were a threshold and/or rounding. Many of the NSIs in our review used variants of small cell adjustment, in that small cells were rounded to, for example 0 or 3, removing all cell values of 1 and 2 from appearing in outputs. This included Norway [6], Estonia [7], Australia [8], New Zealand [9], and the Netherlands [10].

The US census bureau is well known for applying differential privacy, particularly on the 2020 census tables [11]. Although a small number of other statistical products like on-the-map also use differential privacy, and the approach is due to be expanded to other products, at the time of writing the majority of non-census outputs are still protected by non-differentially private methods, namely a threshold (suppression) and table redesign.

(Nordholt and others 2024) [4] is a paper from an EU commissioned working group on disclosure control. Various SDC methods are compared and the impact on outputs, in terms of protection, impact on additivity and consistency are examined. In

a survey of EU member states, record swapping is less commonly used than perturbation, although the reasons for this are unclear. Additionally we are not aware of any application of record swapping to administrative data for official statistics. A combination of targeted record swapping and cell key perturbation remains the recommended approach by Eurostat for demographic data [5].

Updates to cell key perturbation

Cell key Perturbation adds “noise” to the outputs, which makes small changes to the values; for example, adding noise of plus one so a count of four appears as a five. This is particularly effective at protecting against “differencing”, where multiple datasets are combined or “differenced”, such that users cannot be sure whether small differences represent a real person or are caused by the perturbation. When a difference between two cell values is calculated, this number will be the real difference between the values, plus two sets of noise from perturbation. Cell key perturbation was applied to Census 2021 outputs which is described by (Spicer 2019) in more detail.

Several updates have been made to the implementation of cell key perturbation from Census 2021, the most significant of which is changes to the noise added. For Census 2021 outputs, the noise added was from a bespoke distribution, approximately normal shaped centred around a mean of zero. For future iterations, the noise will be approximately Laplace shaped, in line with international best practice. For further information, see Abowd (2023) ‘Confidentiality Protection in the 2020 US Census of Population and Housing, Annual Review of Statistics and Its Application’. [13]

To enable the use of cell keys, and the consistency they provide, this Laplace noise is stored within the ‘ptable’ file – or perturbation look-up file. The ptable for census 2021 used 256 cell keys, allowing 256 different values of noise for each cell value. Each cell key contains a perturbation value, which could either leave the cell unchanged, reduce the cell value by one, increase the cell value by one, and so on. By changing one key, we can change the rate of perturbation only in plus or minus 0.4% increments.

Future uses of cell key perturbation will be using an increased range of 0 to 4,095 which allows a choice of more precise rates of perturbation, in plus or minus 0.02% increments. It also allows large perturbations to be applied with low probabilities, introducing the possibility of large changes in cell values, although keeping these changes infrequent to preserve utility. Any range of record keys can be used for perturbation, though larger ranges increase the file size of the ptable (proportionally). This increased precision allows the ptable to more closely follow the Laplace distribution. The increased range of record keys allows the possibility of applying larger perturbations, with probabilities less than 0.4%. Unlike a theoretical Laplace distribution however, the ptable is still limited in size and does not allow the possibility of perturbing by large amounts, e.g. noise addition of ±20, which would happen very rarely.

Census 2021 used a higher rate of perturbation for ‘small’ cell values’, and a lower rate of perturbation for ‘large’ cell values, to reflect the relative risk of these cell sizes. The same approach will be taken for FPMS, but the rates of perturbation have been smoothed, so that the rate of perturbation gradually reduces as cell sizes are larger, rather than large changes occurring at particular thresholds. Intuitively, this is treating larger counts as having progressively lower disclosure risk, avoiding a ‘cliff-edge’ where cell counts slightly below a threshold have very different perturbation rates than cell counts slightly above a threshold.

Producing Record keys based on ID numbers

To apply cell key perturbation, a set of record keys are required, one for each record in the dataset. These keys need to be random, uniformly distributed integers. As mentioned, the range of the record keys will now be 0 to 4095 allowing more precise control over perturbation rates and distribution.

The record keys need to remain consistent over time, changes in the record keys will result in changes to the outcome of perturbation. The Demographic Index is a spine of data used in the creation of the ‘admin-based census in development’ and contains unique integer ID numbers for each record. Our intention is to base the record keys on the Demographic Index ID numbers, by taking modulo 4096.

The benefits of using ID numbers as the record keys are mostly practical, in that the ID numbers will remain consistent over time and well maintained. New ID numbers are generated sequentially when records are added to the Demographic Index, but the ID numbers are not meaningful of themselves or indicative of any attribute values. The modulo of the ID numbers in the existing admin-based census in development is closely uniformly distributed, making them suitable for use with a ptable.

ID number	Record key Mod(ID number, 4096)
400008828	1660
400010901	3733
400000494	1518
400007157	4085
400006355	3283
400008804	1636
400002869	3893
400001229	2253
400011412	148
400010350	3182

Table 1: example ID numbers and corresponding record key values

Longitudinal differencing

The use of administrative data to produce statistics will enable more timely and frequent outputs. This introduces a new risk arising from longitudinal data: small differences in counts over time.

As an example, in June 2024, there is one person with particular attributes in an area. Then, in a reference period of June 2025, there are two, which reveals that one new person has moved into the area. Or, if an intruder knows there is one new arrival, this reveals their attributes.

June 2024	Ethnic group				
Highest level of qualification	Asian	Black	Mixed	White	Other
Level 1	0	2	33	**286**	0
Level 2	4	1	**1**	**135**	0
Level 3	3	0	6	**251**	4
Level 4+	**12**	2	11	**424**	3
Other	1	0	5	**122**	2

June 2025	Ethnic group				
Highest level of qualification	Asian	Black	Mixed	White	Other
Level 1	0	2	33	**291**	0
Level 2	4	1	**2**	**137**	0
Level 3	3	0	6	**249**	4
Level 4+	**14**	2	11	**420**	3
Other	1	0	5	**126**	2

Figure 1: A hypothetical set of tables of ethnic group by highest level of qualification, 1 year apart. Small changes can be observed in the Asian and Mixed ethnic groups

June 2024	Ethnic group				
Highest level of qualification	Asian	Black	Mixed	White	Other
Level 1	21	12	78	140	1
Level 2	4	7	3	198	1
Level 3	7	4	9	269	2
Level 4+	32	17	115	395	3
Other	2	5	5	160	2

June 2025	Ethnic group				
Highest level of qualification	Asian	Black	Mixed	White	Other
Level 1	21	**13**	**82**	**138**	**2**
Level 2	**5**	**6**	**4**	**197**	**0**
Level 3	**9**	4	**11**	**265**	4
Level 4+	**30**	**20**	**111**	**384**	1
Other	**3**	5	**6**	**171**	2

Figure 2: A hypothetical set of tables of ethnic group by highest level of qualification, 1 year apart. Changes in cell values are much more common, and would be hard to attribute to an individual

Whether this risk requires specific protection depends on what proportion of cells would be expected to change between releases. If a high proportion of cells were to remain unchanged between two releases, it could be inferred that small changes are likely to represent individuals moving to or from an area, potentially allowing an identification and disclosure of their attributes.

Figure 1 shows a hypothetical scenario, breakdowns of ethnic group by highest level of qualification for a given LSOA, for two consecutive time periods. Cells with differences between the time periods are highlighted. In this scenario the majority of cells are unchanged, with some notable differences including one additional mixed ethnic group resident appearing prior to June 2025. Assuming no protection from swapping or cell key perturbation, these tables could reveal that a mixed ethnic group resident has migrated into the area, and their highest level of qualification is level 2.

Some attributes would not be expected to change over time, such as age (date of birth), sex, ethnic group. Other attributes will change over time, such as location, occupation, marital status. Changes in ‘static attributes’ will be much more noticeable in the data, though it is possible that changes in static attributes could be used to disclose information on any other attribute.

As an initial assessment of the stability of estimates in FPMS figures, changes between 2021 and 2022 were observed from the Dynamic Population Model [14], estimates of single year of age by sex at Local Authority level. The average absolute change in cells in these tables was 29. For example, in the Ashfield Local Authority, the number of 20 year old males in 2022 was 37 lower than the number of 19 year

old males in 2021 (all residents will have aged by one year and moved into a higher age category, assuming the age is calculated on the same day of the year). At Local Authority level, these natural changes in population are sufficient to mask movements made by individuals. Broken down into LSOA level changes, these will be much smaller. There are on average 115 LSOAs per Local Authority, so we would expect the changes between years to be much smaller, although there are several factors that make this effect skewed towards some groups, and hard to predict:

- Younger age groups are more mobile, and larger changes would be expected in younger age groups than older
- Migration would be higher in urban areas and areas with mobile populations, especially students, and correspondingly lower in rural areas with more static populations
- These differences between time periods are net differences, some apparent small differences will not represent individuals, but several overlapping inflows and outflows
- Changes are likely to be clustered in larger cell values, changes in smaller counts will be less common and more noticeable

Migration patterns at low geographies could also be observed in Census 2021 data, using resident address one year ago, however this may not represent the patterns that will be seen in administrative data. Capturing movement patterns is more complex in administrative data, with the possibility of linked sources containing different addresses, and related lags in data being present. For example, when residents change address, some data sources may be updated quickly, while others like GP registers could be updated much later.

Further investigation will need to be carried out on FPMS data at LSOA level, to observe the overall impact of these overlapping factors.

Protecting against longitudinal differencing

Cell key perturbation is designed to protect against disclosure by differencing. However, it is also designed to ensure consistency between cells that contain the same records. If the same four records, say, appear together in different cells, they will have the same record keys, and therefore receive the same perturbation. For cell values that are static over time, the same perturbation value will be added in every time period. Differences in counts over time will never be caused by perturbation, and therefore the differences would be known to represent genuine changes.

If longitudinal differences are deemed to be a significant disclosure risk, the cell key perturbation could be amended to protect against them. By using different record keys for each time period, the perturbation for data for June 2025, for example, would be independent of that for June 2024. The same four records appearing together would have different record keys and could receive different values of perturbation. This would mean that where changes over time can be observed in cell values, there is sufficient doubt whether the changes represent real individuals or were introduced by perturbation.

Targeted Record Swapping

Targeted record swapping was the main form of protection for both 2011 Census and Census 2021. Households considered most at risk of identification or disclosure are swapped with “matched” households in another geographical area by swapping their geographical information. The matched households are chosen to be similar on basic attributes, such as the number of people in a household, and most swaps are performed within a local authority.

In this way, considerable doubt is introduced to smaller counts at low geographies, while higher level information at the local authority level is mostly unaffected. Another benefit of this approach is that swapping can be targeted to protect specific risks, for example targeting low marginal counts on certain variables which pose a greater risk in a flexible outputs environment.

A combination of targeted record swapping and cell key perturbation are the recommended disclosure control methods by Eurostat for demographic data [5]. Targeted record swapping will target individuals assessed as at risk of identification or disclosure. Cell key perturbation will introduce uncertainty designed to protect against differencing.

As with perturbation, it’s important to consider how the features of FPMS data could impact the swapping, and the protection it provides.

Sets of outputs will be produced annually, although it is possible that future outputs will become more frequent where the data allow this and there is a user need. If outputs are more frequent, it may be more difficult to apply the swapping process in a timely manner. A simplified swapping process may be needed to achieve this.

The impact of swapping over time will also need to be considered. One aspect is ensuring that by comparing several sets of outputs, it does not become obvious where swapping has been applied. For example, if an individual is targeted for swapping in 2023, but not in 2021, 2022, or 2024, their absence and reappearance in their LSOA may be conspicuous and draw attention to the protection method. This would also have unintended consequences if several sets of swapped data were used for migration analysis, this individual would erroneously appear to have moved address in 2023 and then moved back. If post-swapping geographies were used naively to study migration patterns, the scale of internal migration would be significantly over-estimated.

One other factor to consider for FPMS data, is how additional data could be appended by linkage long after the ‘collection’ of the data. It is likely that to some extent, more data or topics will become available retrospectively for a given release. For example, the ONS might produce a dataset of all residents as of June 2023, containing age, sex, ethnic group at LSOA level, then at a later date, more data on other topics becomes available via linkage. ONS needs to consider the likely disclosure risk of the final form of the data, and if possible, protect against future disclosure control risks. Failure to consider and protect against future risks may inhibit our ability to utilise the data to the fullest extent.

In our review of other NSIs, we did not find examples of record swapping being applied to administrative data. If our pilot application of swapping does not provide good disclosure protection or is not possible due to the structure of the data, then alternative methods will need to be applied instead. This will most likely be an increased rate of perturbation, with added protection of small values.

FPMS data will be available at LSOA level

The current implementation of the admin-based census has geography information available at LSOA level. This reduced geographic detail relative to census will in general lower the disclosure risk, being harder to identify an individual in an LSOA (~1500 residents) rather than Output Area (OA ~300 residents). This also has an impact on our targeted swapping. For census 2021, households were targeted if they contained residents who were unique on an attribute, for example, they were the only widow within the OA, or they were the only person of mixed ethnic group. At the larger LSOA level, and, given the reduced granularity and number of other attributes, far fewer residents are unique. Targeting fewer records for swapping naturally reduces the protection that it provides. To ensure a sufficient rate of swapping, the criteria for risky records may be broadened to include more records, or the targeted swapping may be accompanied by a level of random record swapping. Alternatively, it may be that only a reduced level of swapping is needed at LSOA level as the records will be less identifiable within a larger geography.

FPMS data is not based on households

Unlike surveys and the census, most linked administrative datasets are address based and will at least initially not contain household information at the record level. The targeted record swapping was previously applied at the household level such that households containing risky records were swapped with each other. Without the household level information, this approach will not be possible. The obvious alternative is to swap all residents within an address with all residents in another address, in place of household. This may not work as intended particularly with larger addresses that would previously be considered multiple households. Changing information on larger addresses also has an outsized impact on the utility of the method as a large number of residents would be swapped collectively. The other alternative is to swap individuals with individuals in other areas. This would lead to less swapping overall but there are utility reasons to keep residents of a household together, to preserve clustering and other patterns that often occur within households.

Transparency

As with Census 2021, in the interest of transparency, ONS intends to publish details of our disclosure protection including an average rate of perturbation, and a range of rates of record swapping. It will also be known that the noise distribution used in perturbation will be close to the Laplace distribution. This transparency is intended to aid analysis, and inform users of the likely impact the disclosure control methods will have on their findings. Our updates to cell key perturbation mean that the rate of perturbation, the proportion of cells that receive non-zero noise, will be highly dependent on the cell values. Reporting an accurate ‘average’ may be difficult as the

amount of noise added to any frequency table will depend on the size of the cells within that table. In this way two different tables produced by an analyst, say one at the National level and one at LSOA, may have very different perturbation rates.

Differential Privacy Suitability

An algorithm M is differentially private if any two datasets X and X', which differ by one record, have very similar probabilistic outcomes (the ratio of probabilities of any outcome is less than e^{ε}, where epsilon (ε) is the primary parameter of the method). Algebraically, an algorithm M is differentially private if: $p(m(x) \in S) <= p(m(x') \in S) * e^{\varepsilon}$. (Dwork and Roth 2014 [15]). Where X and X' are datasets differing by one record. Differential privacy is a relatively new standard of protecting data, which has received much interest and attention in recent years. The US Census Bureau used the differential privacy standard to protect outputs from their 2020 census, several large technology companies also have differential privacy implementations, including Google [16] and Apple [17].

The standard form of the method is controlled using the epsilon value, using low values for epsilon provides strong protection against disclosure risk. However, a strict implementation without considerable 'post-processing' of results would likely impact the usefulness of data released, potentially impacting the effectiveness of research carried out. The epsilon parameter determines the strength of the 'privacy guarantee' of differential privacy. In practice, using a Laplace noise mechanism, the epsilon parameter determines how much noise is added to cell values. Low epsilon values result in more noise being added, and more protection to the data. Higher epsilon values result in less noise being added, so less protection is added but the impact on utility is lower. Supporters of differential privacy argue that the balance between risk and utility can be directly controlled through the choice of epsilon, and have applied cost benefit analysis to determine their parameter selection. See [18] for information on how the US Census Bureau set their epsilon parameter.

An important concept in differential privacy is the allocated 'privacy budget', measured by the epsilon value. Protecting a single output with epsilon differential privacy would have a 'privacy budget' of epsilon. Producing multiple outputs makes the privacy budget more complicated, each independent output adds to the disclosure risk and affects the privacy budget. Since FPMS data will be available to many sets of analysts with independent projects, it may be difficult to stay within a fixed privacy budget. As the data will be used for several years at least, it will be impossible to know in advance all of the outputs that will be produced, and set a privacy budget accordingly. The impact of this variety of outputs and users over time is one major reason it would be difficult to stay within a fixed privacy budget, thereby adhering to the formal definition of differential privacy.

ONS has carried out a previous pilot study of differential privacy on mortality data (Dove 2019 [19]) which took a similar approach to the US census bureau to create a set of microdata that was differentially private, as opposed to a list of frequency tables. This demonstrated the potential of the approach but highlighted practical

issues that need to be considered, particularly in how noise was added to cell counts of zero. In the pilot, negative cell values arose from negative noise being added to small cell values including zeros. The adjustment of these negative counts was a major source of bias towards less common categories, particularly from zero value cells, which can form a considerable proportion of cells in sparse tables.

Differences between cell key perturbation and differential privacy:

- Zero perturbation - standard cell key perturbation does not add noise to cells of zero. The ONS has developed an extension of perturbation that does perturb cell counts of zero, however, not all zeros have the possibility of being perturbed. The method avoids perturbing structural zeros, and combinations not observed at a higher geography, indicating an unusual combination of attributes. Overall, zero cells receive less noise and are less likely to be perturbed than non-zero cells.
- Non negative counts – small cell values do not receive negative noise larger than themselves, for example a cell value of two cannot receive noise of minus three or less. This avoids apparent negative counts appearing in public outputs, or the need for adjusting negative counts which can inadvertently introduce biases.
- Capped sizes of perturbation – cell key perturbation only applies a limited range of noise, which introduces cases that cannot meet the e^{ε} ratio of probabilities. Relaxed versions of differential privacy are being developed (epsilon-delta DP) although exceeding the ratio is recommended to only happen very rarely.
- Privacy budget – outputs from a differentially private system are typically limited, as further outputs are considered to reveal more information and reduce the protection. ONS will not be able to determine all of the outputs ahead of time, particularly considering the use of our data in secure environments by external users.

Epsilon-Delta differential privacy

Our application of Cell key perturbation uses limited values of noise stored in a look up file, known as the ptable. Unlike the Laplace distribution, it does not contain the possibility of large values of noise being applied, albeit with small probabilities. In these cases, the ratio of probabilities between outcomes of two similar datasets is unbounded (greater than e^{ε}) hence the result is not differentially private.

Some measures of differential privacy allow the ratio of e^{ε} to be exceeded with a given probability, delta [15], e.g. for an epsilon-delta mechanism with epsilon of 2 and delta of 0.01, the probabilistic outcomes of two datasets X and X' are bounded by e^2, with a probability of 0.99. There is a 1% probability that the ratio is greater than e^2.

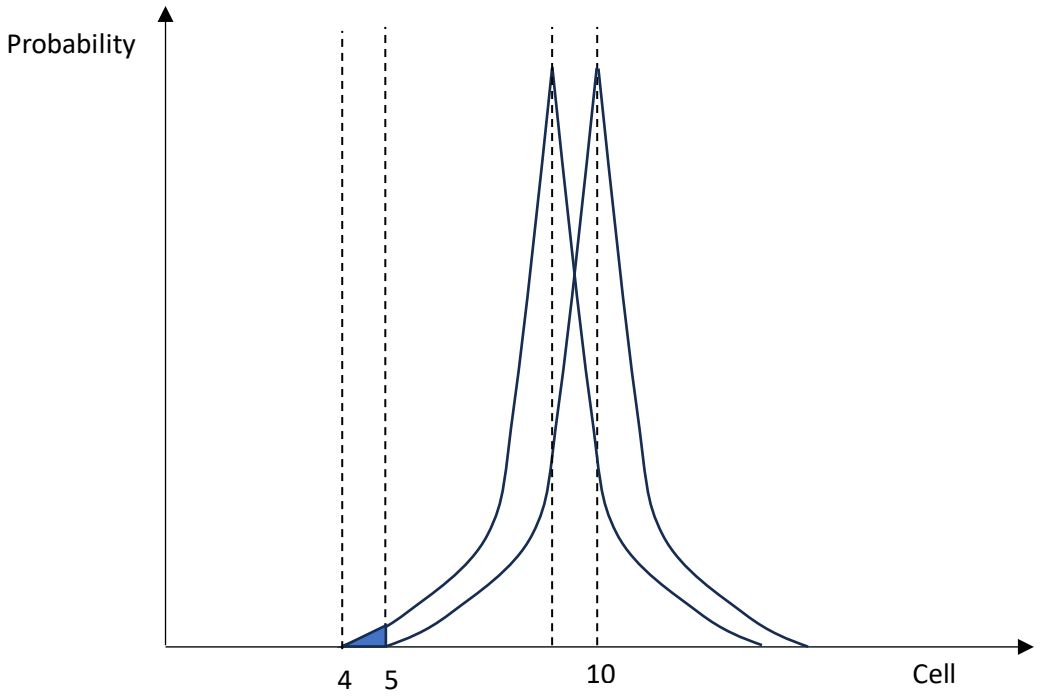

Figure 3: Illustration of truncated noise applied to two datasets differing by one record. Since the distribution is not asymptotic, the ratio of the two curves is not bounded on the edges (highlighted).

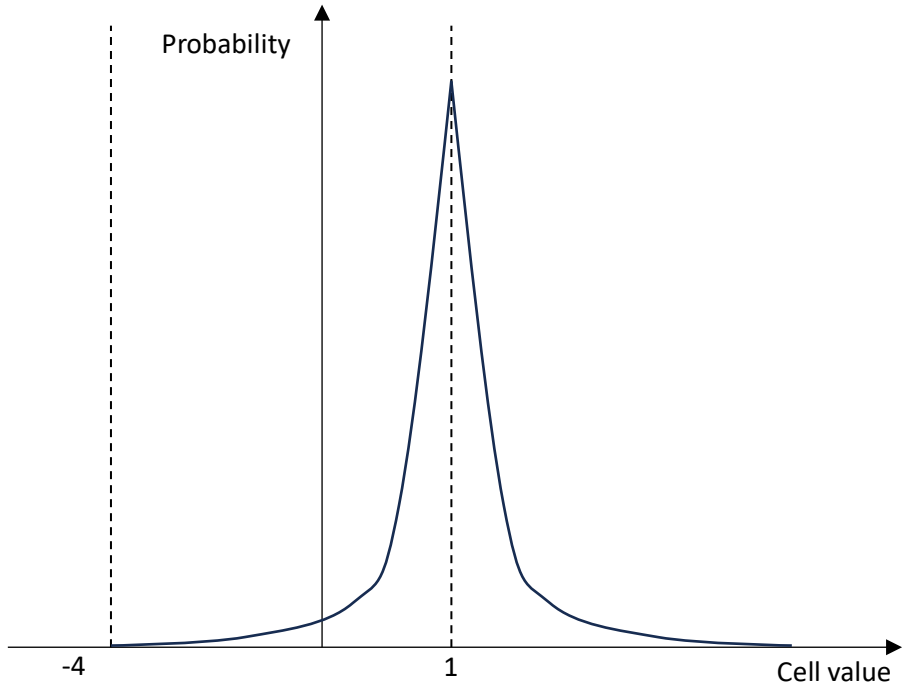

Figure 4: Standard Laplace noise applied to cell values of 1 would result in negative counts, which would require adjustment. The noise that cell key perturbation applies to cell values of 1 is not symmetric, to avoid negative values.

The ratio between the two curves is unbounded where p(X'=x)=0 but p(X=x)>0, or vice versa. In this example illustrated in figure 3, a given cell value (say from dataset

X) before perturbation is 10. The perturbation provides probabilities that the observed cell value could be between 5 and 15. A dataset X' with one record removed has potential values after perturbation of 4 to 14. In the event that the observed cell value was 4, it would be known with certainty that the dataset producing this value was X' (in which case the datasets X and X' were not indistinguishable).

For an overall system of cell key perturbation this will be difficult to measure, as smaller cells receive more noise than larger cells. Measuring the epsilon and delta values for small cell values will indicate more strict privacy guarantees than for larger cells.

ONS will undertake another application of Differential privacy, including an amended noise distribution intended to avoid the bias issues. ONS will undertake research on the eligibility of our current methods, particularly cell key perturbation, to be considered differentially private, including the epsilon-delta variant. It will also investigate how multiple independent outputs can be considered when calculating a privacy budget.

Our intended methods for FPMS are targeted record swapping, and cell key perturbation. The updated perturbation applies Laplace noise to frequency tables, a very common approach to ensuring differential privacy (Dwork and others 2006 [20]).

ONS does not claim that our disclosure methods will be differentially private due to the differences listed above which mean that the definition of epsilon Differential Privacy is not met in all circumstances. The cell key method does possess strong similarities to many DP approaches in that Laplace noise is added to frequency tables, and zero cells are also perturbed in our application. It is possible that this approach meets the conditions of epsilon-delta differential privacy. Further research will be carried out to investigate whether the conditions are met, and how a set of outputs protected by perturbation fit within a privacy budget.

Next Steps

- Carry out another application of differential privacy, including alternative applications of noise, and research into whether cell-key perturbation meets the criteria for epsilon-delta differential privacy
- Attempt a pilot application of record swapping on a realistic data model of the linked administrative data, and ensure our process will be compatible with existing pipelines and data structures
- Discuss the anticipated content and frequency of outputs, including the content of the recommendation, which will determine what disclosure control is required
- Decide a rate of perturbation based on the expected scope and frequency of outputs, and considering whether the data will have protection from swapping

References

[1] Dove I., 2023 ‘Disclosure control proposal for Future Population and Migration Statistics’, Office for National Statistics

[2] Office for Statistics regulation, 2022 ‘UK Statistics Authority's Code of Practice for Statistics’

[3] Hundepool, A. and others, 2012. ‘Statistical Disclosure Control’ Wiley Series in Survey Methodology

[4] Nordholt and others, 2024 ‘Guidelines for SDC methods for Census and Demographics Data’. Methods and Tools for Time Series, Seasonal Adjustment and Statistical Disclosure Control

[5] Recommendations for the protection of Census data, Collaboration in Research and methodology for Official Statistics

[6] Personal data in the statistics – SSB, Statistics Norway

[7] Tiit E.M., 2014 2011_Population_and_Housing_Census._Methodology.pdf (stat.ee) Statistics Estonia

[8] ‘Confidentiality and relative standard error’ Australian Bureau of Statistics, 2021

[9] ‘Methodological standard for confidentiality in the 2023 Census Statistics New Zealand, 2023

[10] Hundepool A. and De-Wolf P, 2010 ‘Statistical Security’ Central Bureau of Statistics

[11] Differential Privacy and the 2020 Census, US Census Bureau, 2021

[12] Spicer, K., 2019 Statistical Disclosure Control SDC for 2021 UK Census

[13] Abowd, J (2023) ‘Confidentiality Protection in the 2020 US Census of Population and Housing, Annual Review of Statistics and Its Application

[14] Blackwell, L. 2022 Dynamic population model for England and Wales

[15] Dwork C. and Roth A, 2014 ‘The Algorithmic Foundations of Differential Privacy’ Foundations and Trends in Theoretical Computer Science

[16] Guevara, M 2021 ‘How we’re helping developers with differential privacy’ Google

[17] Differential_Privacy_Overview Apple

[18] Key Parameters Set to Protect Privacy in 2020 Census Results. US Census Bureau press release, 2021

[19] Dove, I. 2019, Applying differential privacy protection to ONS mortality data, pilot study

[20] Dwork C., Mcsherry F., Nissim, K., Smith, A., 2006 Calibrating Noise to Sensitivity in Private Data Analysis Theory of Cryptography

[21] Rinott, Y., O’Keefe, C., Shlomo, N., and Skinner, C.(2018) Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. *Statistical Sciences*, Vol. 33 (3), 358-385.