Comparison of post-tabular statistical disclosure control methods

This paper will evaluate two post-tabular disclosure control methods: rounding plus a threshold and cell key perturbation.

Contact:
Statistical Disclosure Control team
sdc.queries@ons.gov.uk
+44 1329 444789

Release date:
3 May 2024

Next release:
To be announced

Notice

22 November 2024

This paper has been edited. The sentence reading "Average noise, total noise, the proportion of overall counts changed, Hellinger's distance, and robust eigenvalues were lower after the use of perturbation (with and without threshold) compared with the 10-5 rule." was changed to read "Average noise, total noise and the percentage of overall counts changed were lower after the use of perturbation (with and without threshold) compared with the 10-5 rule." as this is more accurate.

Table of contents

1. Overview of statistical disclosure control

2. Benefits of perturbation

3. Benefits of 10-5

4. Methods

5. Impact of perturbation and rounding plus threshold on noise

6. Impact of perturbation and rounding plus threshold on utility

7. Conclusion

8. Cite this methodology

1 . Overview of statistical disclosure control

Statistical disclosure methods are applied to the statistics and outputs produced by us, at the Office for National Statistics (ONS), to protect the confidentiality of data subjects and ensure adherence to ethical and legal commitments to protect data privacy. These methods reduce the utility of the data, so it is necessary to balance the risk of disclosure with the loss of utility.

This methodology working paper will evaluate two post-tabular methods: rounding plus threshold (known as "the 10-5 rule"), and cell key perturbation, both of which were used for census outputs, and will be available in the Integrated data service (IDS). This paper will compare the two methods, particularly their impact on data utility. The analysis may be useful to users of published Census 2021 data, as well as those considering which disclosure control methods to apply.

The 10-5 rule refers to suppressing counts below ten and rounding counts above 10 to the nearest five. Cell key perturbation is a method that adds controlled "noise" to counts in frequency tables. This technique uses an algorithm that applies a pre-defined level of perturbation to cells in each dataset.

Every record within the microdata is assigned a random number (a record key). When frequency tables are constructed, each cell is a count of the number of records, and the cell key is calculated by summing their record keys. The combination of cell value and cell key is then read from a "p table", a look-up file that determines the amount of perturbation that should be used, known as the "p value". This is a positive or negative whole number that will be added to the original count. Where the same combination of records appears in different tables, all instances will have the same cell value and cell key, and so receive the same perturbation.

A similar process is used for the perturbation of cells with a count of zero. A random number is assigned to each category of each variable and used to produce a random and uniformly distributed category cell key, in a similar way to the cell key. This category cell key is used to make a random selection of cells to perturb.

It was found that perturbation produces less noise than 10-5 and provides better utility. For this reason, we generally recommend the use of perturbation over 10-5 for statistical disclosure control. 10-5 may be preferred to keep consistency with previous releases, or where the data are particularly sensitive.

2 . Benefits of perturbation

Cell key perturbation was one of the main forms of statistical disclosure control in Census 2021, described in our Protecting personal data in Census 2021 results methodology. Perturbation has several advantages over the 10-5 rule. It protects data by introducing uncertainty, as users will not know which counts are "real" and which have been affected by perturbation. It also allows small counts to be published, as it will not be known whether these small counts refer to real individuals. Additionally, it protects against differencing, because there will be inconsistencies between the counts when tables are constructed or aggregated in different ways.

Allowing for small counts to be published is an important advantage for frequency tables that will be produced at lower geographies, or with many variables. Including small counts increases the utility of these outputs. This is important because it allows for more accurate and useful data that local authorities and others can use for service provision.

3 . Benefits of 10-5

Despite the increased utility of perturbed outputs, there are several reasons why the 10-5 rule is preferred in some circumstances. For instance, because some outputs were produced before the code for perturbation used for Census 2021 was available as a method, some analysts wanted to continue using 10-5 on Census 2021 outputs for consistency.

An additional benefit of the use of the 10-5 rule is that it is relatively easy to implement because of its simple nature. It requires less time and coding experience than the perturbation method, which is more complex and requires a specific code to be used. It is also clear for analysts and users to see that the 10-5 rule has been applied.

The issues posed by perceived disclosure are another reason for use of the 10-5 rule. For sensitive topic areas that have stronger legal protections, such as ethnicity, religion, and sexual orientation, this is an especially important concern. Perceived disclosure relating to these variables may affect the trust of data respondents if they believe they can identify themselves in the outputs.

We carried out Intruder testing, where "friendly intruders" tried to identify individuals in census data, using the planned outputs, to ensure this was not likely to be done with both confidence and accuracy. According to the United Nations Economic Commission for Europe (UNECE) paper (PDF, 223KB), intruders reported high levels of confidence in their claims, with the average claim being reported with 74% confidence and some claims reported with 100% confidence. Intruders reported these high confidence levels despite knowing that the outputs had been protected with statistical disclosure control methods, including perturbation. This is a particular concern for releasing data regarding sensitive topics, as people will be confident in claims even when they are not likely to be correct, and even when they know disclosure control methods have been applied. For this reason, even where disclosure control methods have been applied, it may also be important to avoid any appearance that unsafe data has been released.

4 . Methods

The analysis compares the use of the cell key perturbation method and the 10-5 rule on the 2011 Census public use file, which is useful as a dataset to demonstrate the effects of these methods because it is openly available. Although analysts will often be using much more complex datasets, this analysis is intended to give an indication of relative utility loss between measures.

For perturbation, we have compared five different settings of the method, applied using five different p tables.

The first of these is the perturbation p table used for Census 2021 with 256 cell keys. The second is an adapted version of the Census 2021 p table, that also uses 256 keys and, in addition to this, suppresses all counts below 10, to represent this method as it will be applied in the Integrated data service (IDS). These were compared with three alternative p tables with higher perturbation rates (meaning they will change a higher proportion of cells in the table) that employ 4,096 keys.

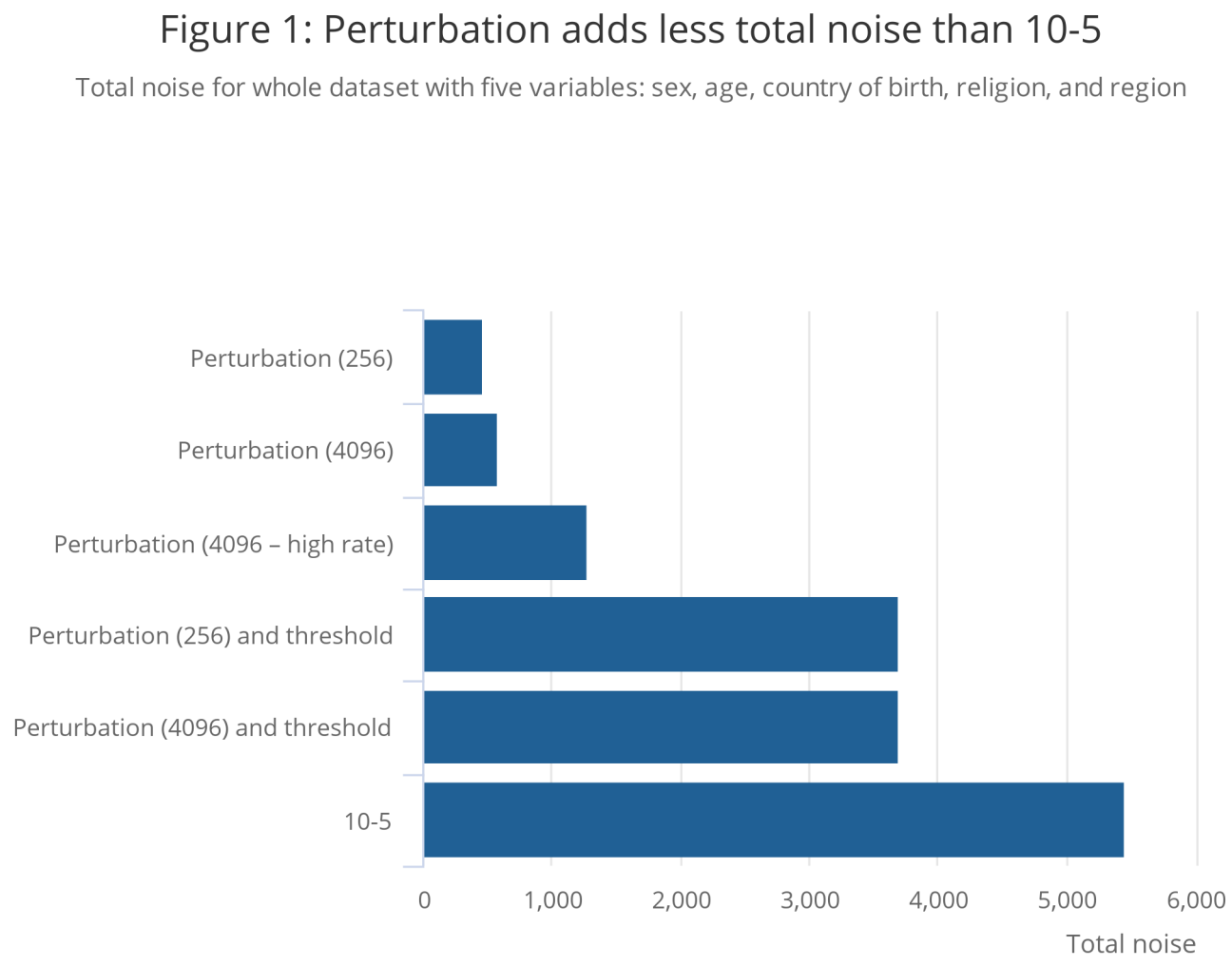
The distribution of p values, such as the noise, in the 4,096 p tables follow a Laplace distribution which has less variance than the distribution in the 256 key p tables. This means that a relatively higher proportion of the p values will be smaller values (for instance, plus one or negative one). These p tables were created with a much larger number of cell keys (4,096 instead of 256) to enable larger p values to be included in the p table. One of these 4,096 p tables suppresses all counts below 10, and another has a higher perturbation rate of Laplace-shaped noise.

5 . Impact of perturbation and rounding plus threshold on noise

In this section, we consider the impact of perturbation, with and without thresholds, compared with the 10-5 rule in terms of the noise added to the dataset. Total noise refers to the total absolute sum of the p values in the final table and average noise refers to the absolute average of all p values in the final table. Proportion of counts changed refers to the proportion of counts from the final table that were changed by the addition of a p value.

Figure 1: Perturbation adds less total noise than 10-5

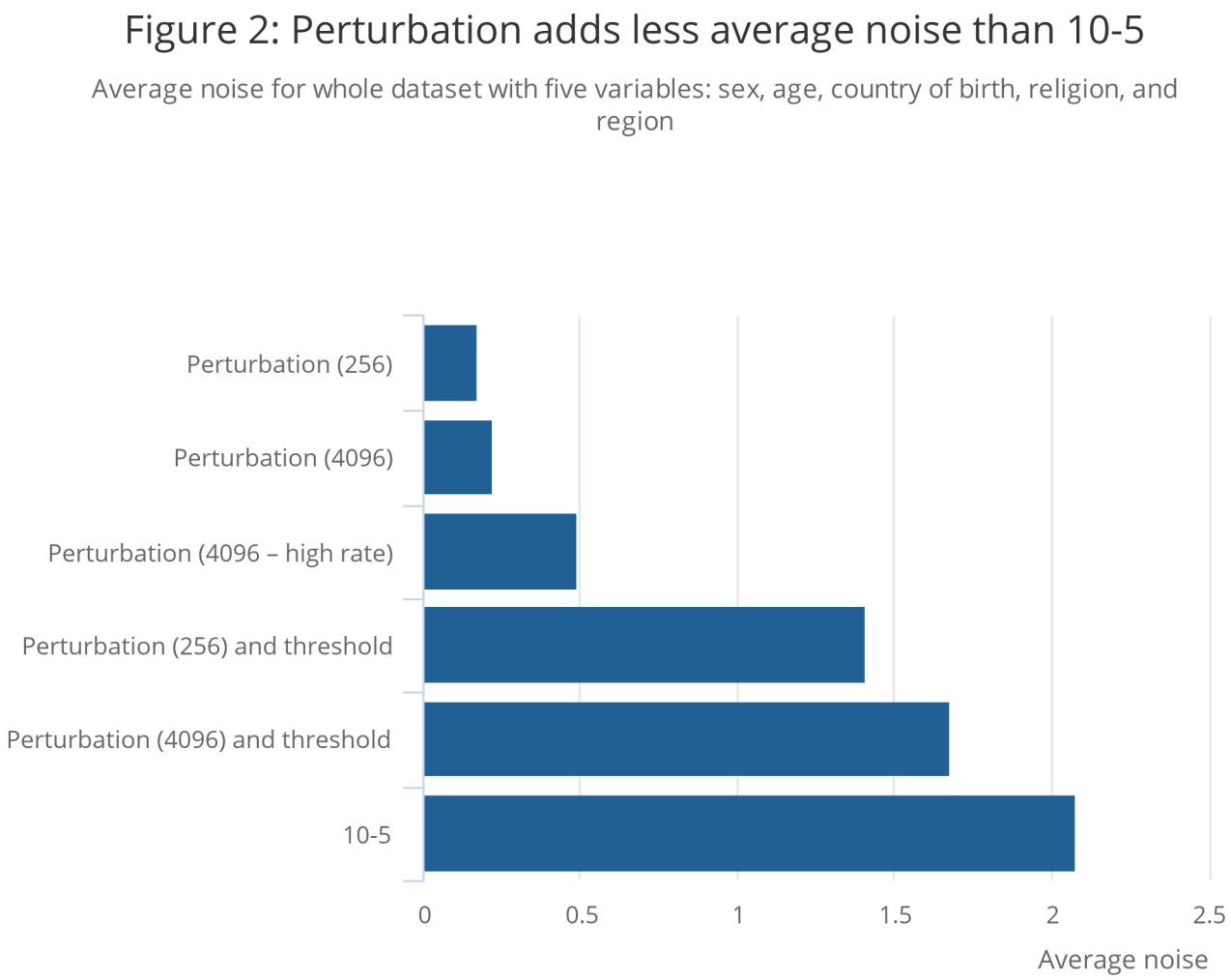
Total noise for whole dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Figure 2: Perturbation adds less average noise than 10-5

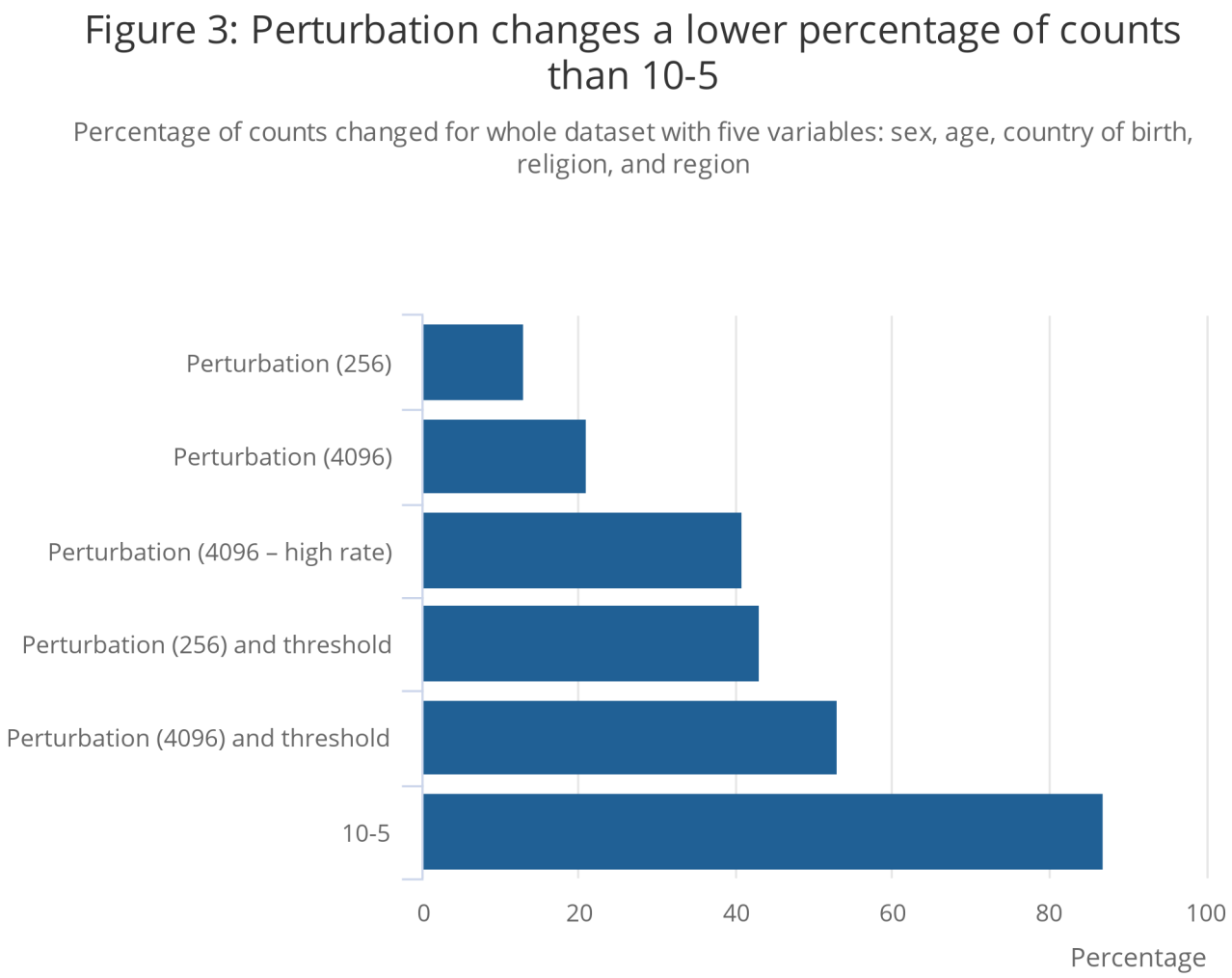
Average noise for whole dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Figure 3: Perturbation changes a lower percentage of counts than 10-5

Percentage of counts changed for whole dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

For a table of five variables using the whole of the teaching file dataset, using the 10-5 rule resulted in much more total noise compared with all types of perturbation. Using 10-5 also resulted in a greater average noise and percentage of counts changed, with over 80% of counts having been changed.

When comparing the different versions of perturbation with threshold, perturbation with 256 cell keys and 4,096 cell keys had similar results, but the 4,096 cell key version resulted in a greater percentage of counts changed.

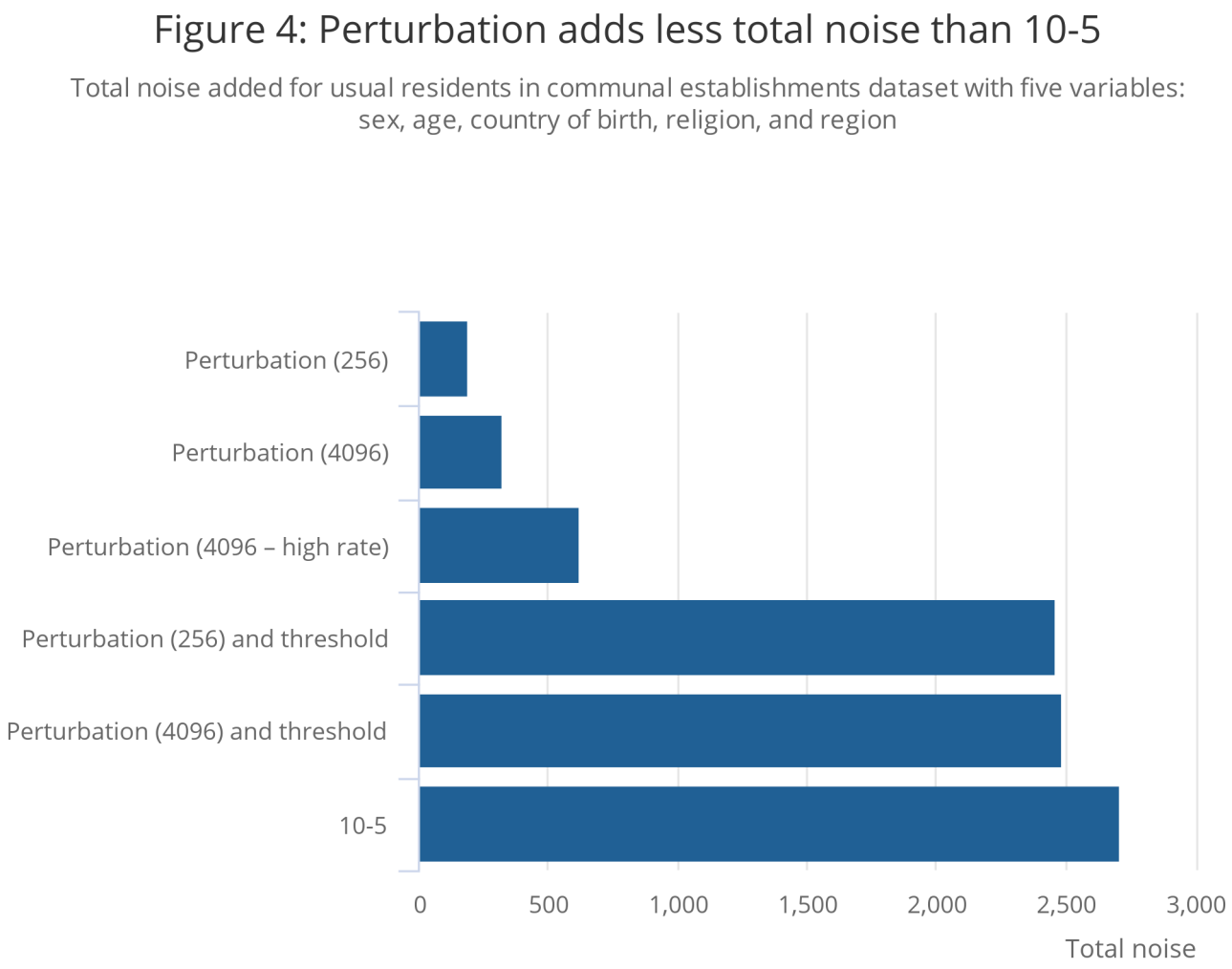
Both 256 and 4,096 key perturbation with threshold resulted in significantly higher total noise, average noise, and percentage of counts changed compared with any perturbation performed without applying a threshold. Applying a threshold will always result in more noise if the table contains small counts below the threshold. Conversely, if the table contains only counts larger than the threshold, applying the threshold will have no impact.

When comparing perturbation without threshold, the 4,096 cell key versions resulted in greater total noise and average noise, and a higher percentage of counts changed, with the high rate 4,096 key version producing the most noise.

To examine the effects of these methods in a very sparse dataset, the same analysis was performed on a subset of the data that includes only residents of communal establishments. On a table with the same five variables used above, very similar results were observed.

Figure 4: Perturbation adds less total noise than 10-5

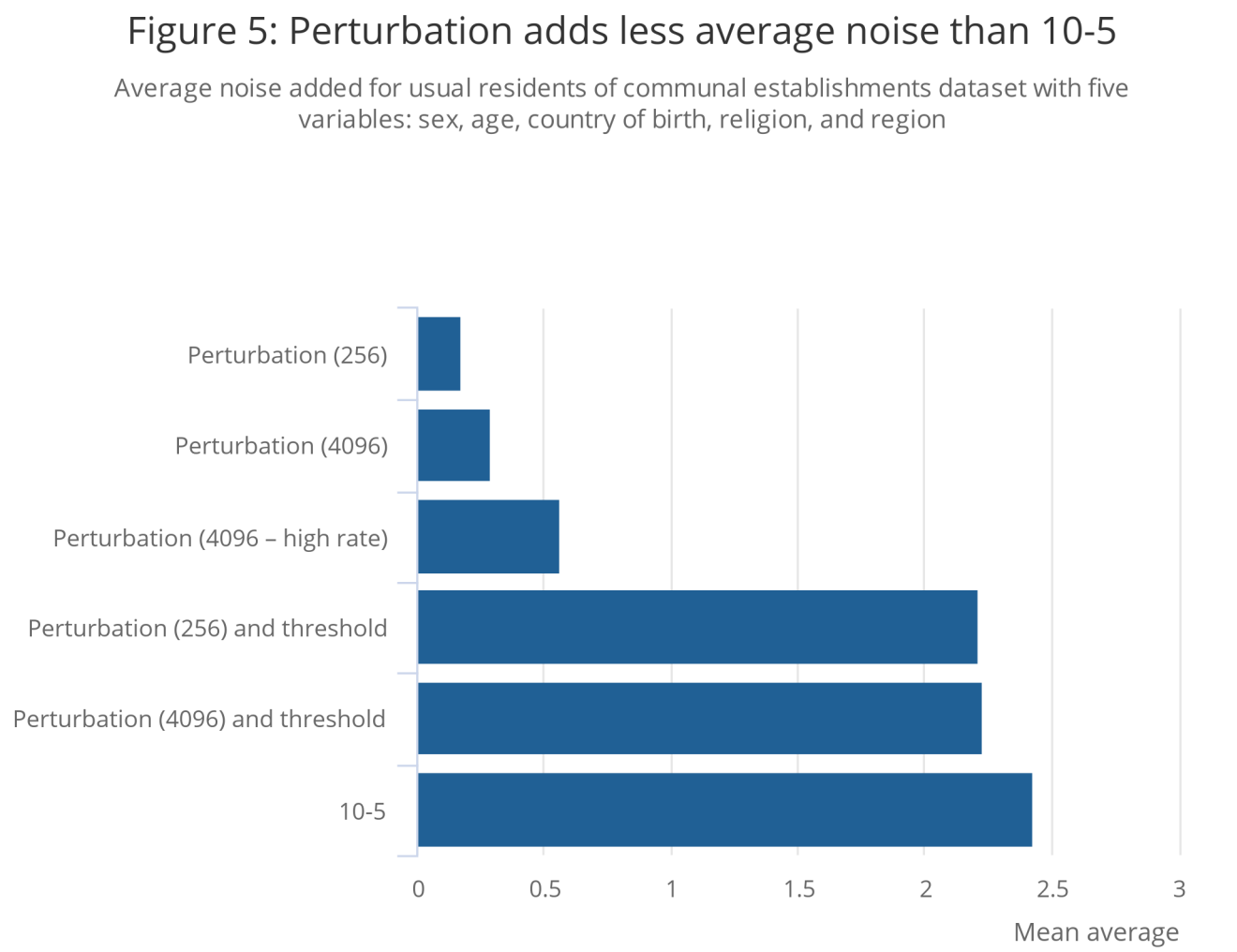
Total noise added for usual residents in communal establishments dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Figure 5: Perturbation adds less average noise than 10-5

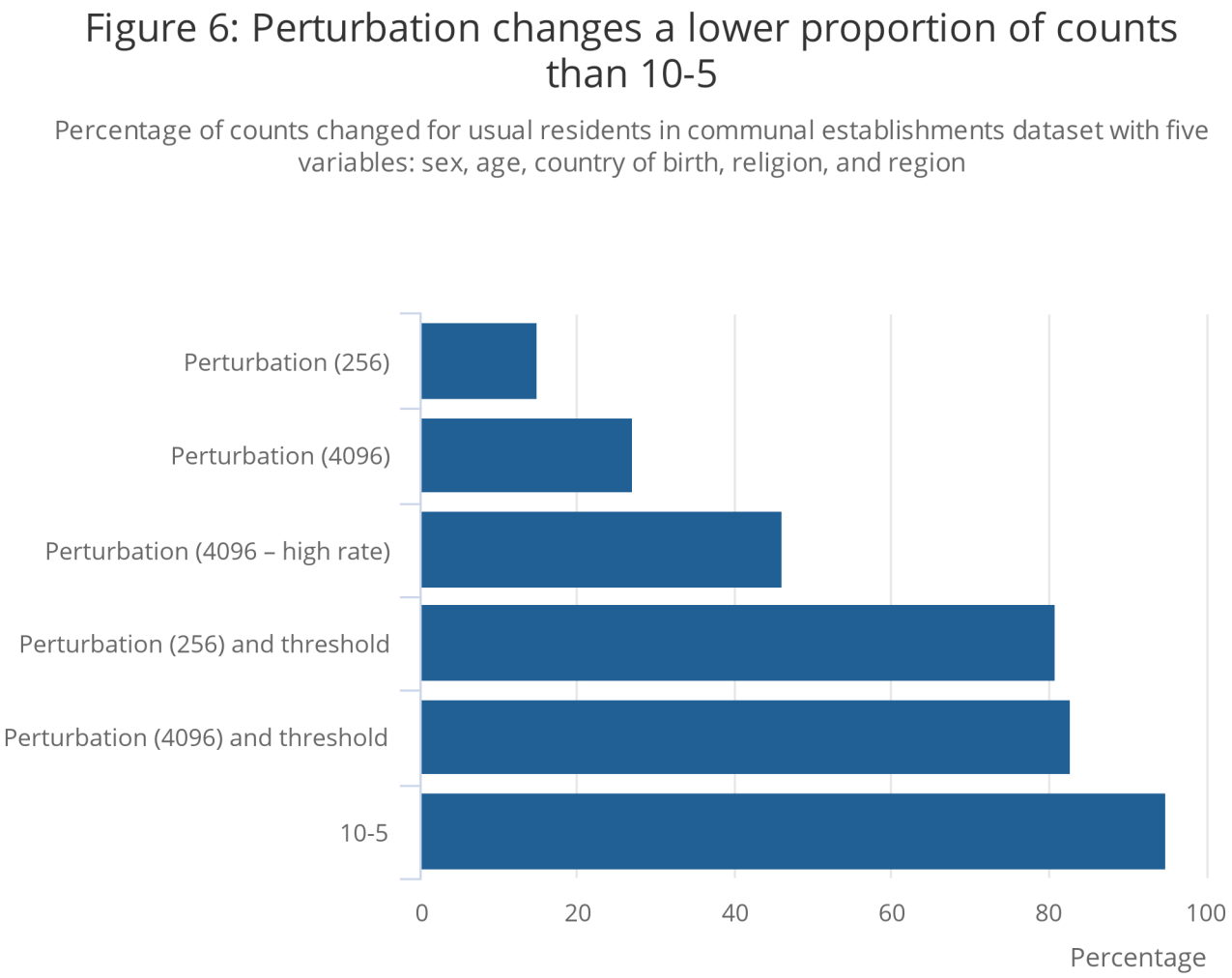
Average noise added for usual residents of communal establishments dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Figure 6: Perturbation changes a lower proportion of counts than 10-5

Percentage of counts changed for usual residents in communal establishments dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Using the 10-5 rule resulted in greater total change, average change, and percentage of counts changed, compared with perturbation with 256 or 4,096 keys. Comparing the two different versions of perturbation without threshold, perturbation with 256 cell keys had a lower total noise, average noise and percentage of counts changed compared with perturbations with 4,096 cell keys.

When comparing the perturbations with threshold, 4,096 key perturbation resulted in slightly higher levels of mean and total change, and a slightly higher percentage of cells changed.

Most counts in this table were under 10, so as expected, all methods that used a threshold (10-5 as well as perturbations with threshold), resulted in a substantially higher total noise, average noise, and percentage of counts changed compared with perturbation without threshold.

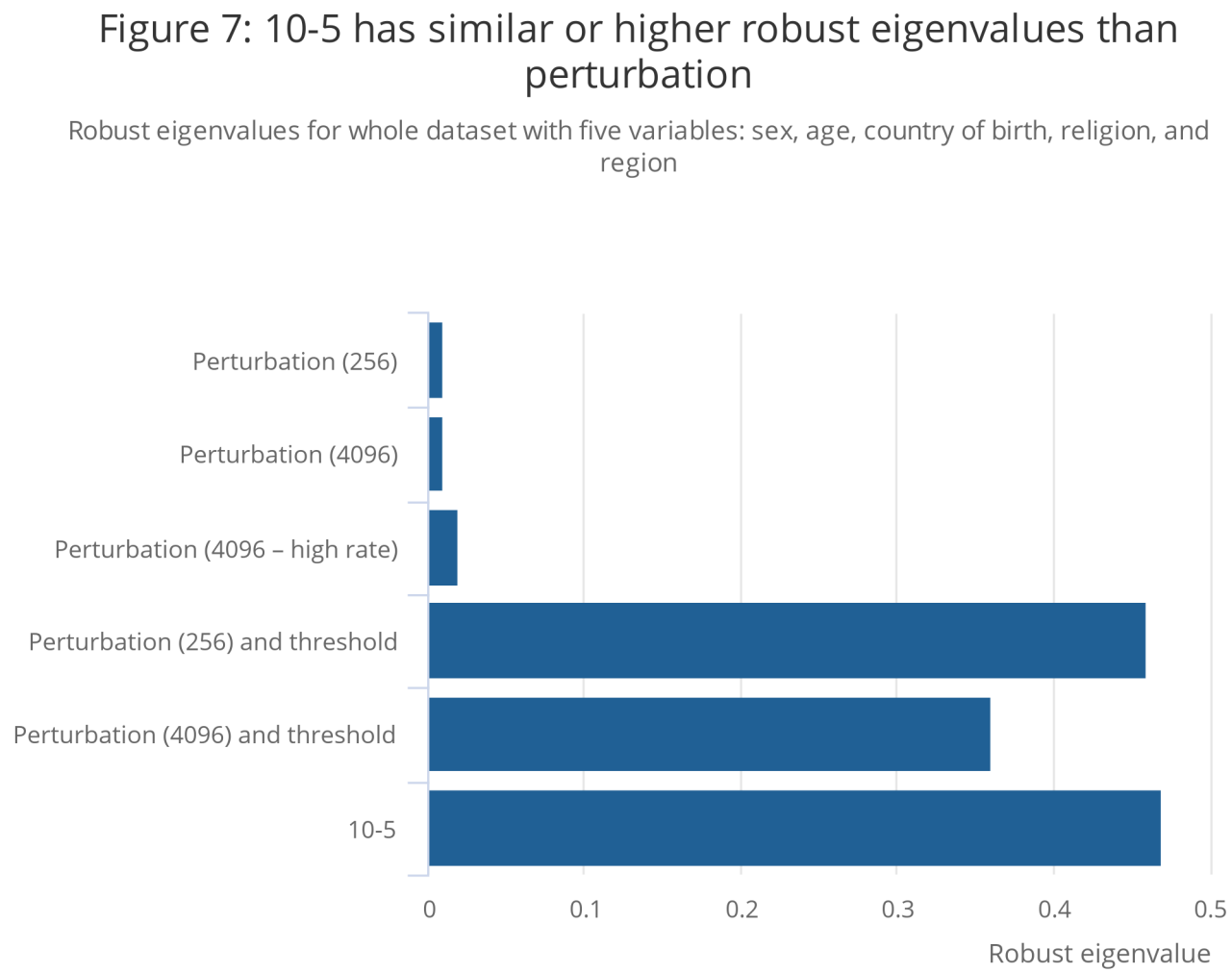
6 . Impact of perturbation and rounding plus threshold on utility

In this section, we consider the impact of perturbation compared with the 10-5 rule in terms of utility. Utility refers to usefulness of the protected data for analysis by researchers. Although the utility of data to a given researcher is very difficult to measure directly, measures of information loss can be calculated as a proxy to compare the impacts of these disclosure control methods.

Hellinger's distance is a distance metric that quantifies the similarity between the original data and the protected data. The lower this value is, the closer the values are, so less information loss has occurred. Hellinger's distance is a commonly used measure of utility in the context of statistical disclosure control. It is useful for comparing statistical disclosure control methods, as shown in the United Nations Economic Commissions for Europe (UNECE) working paper (PDF, 146KB), because it is bounded by zero and approximately the square root of the number of cells in the table, which will be the same across different methods. The Statistical disclosure control (SDC) practice guide defines eigenvalues as a measure that compares the eigenvalues from a robust version of the covariance matrix of the original and protected data. They are also bounded between zero and one, which helps to compare the utility of different methods.

Figure 7: 10-5 has similar or higher robust eigenvalues than perturbation

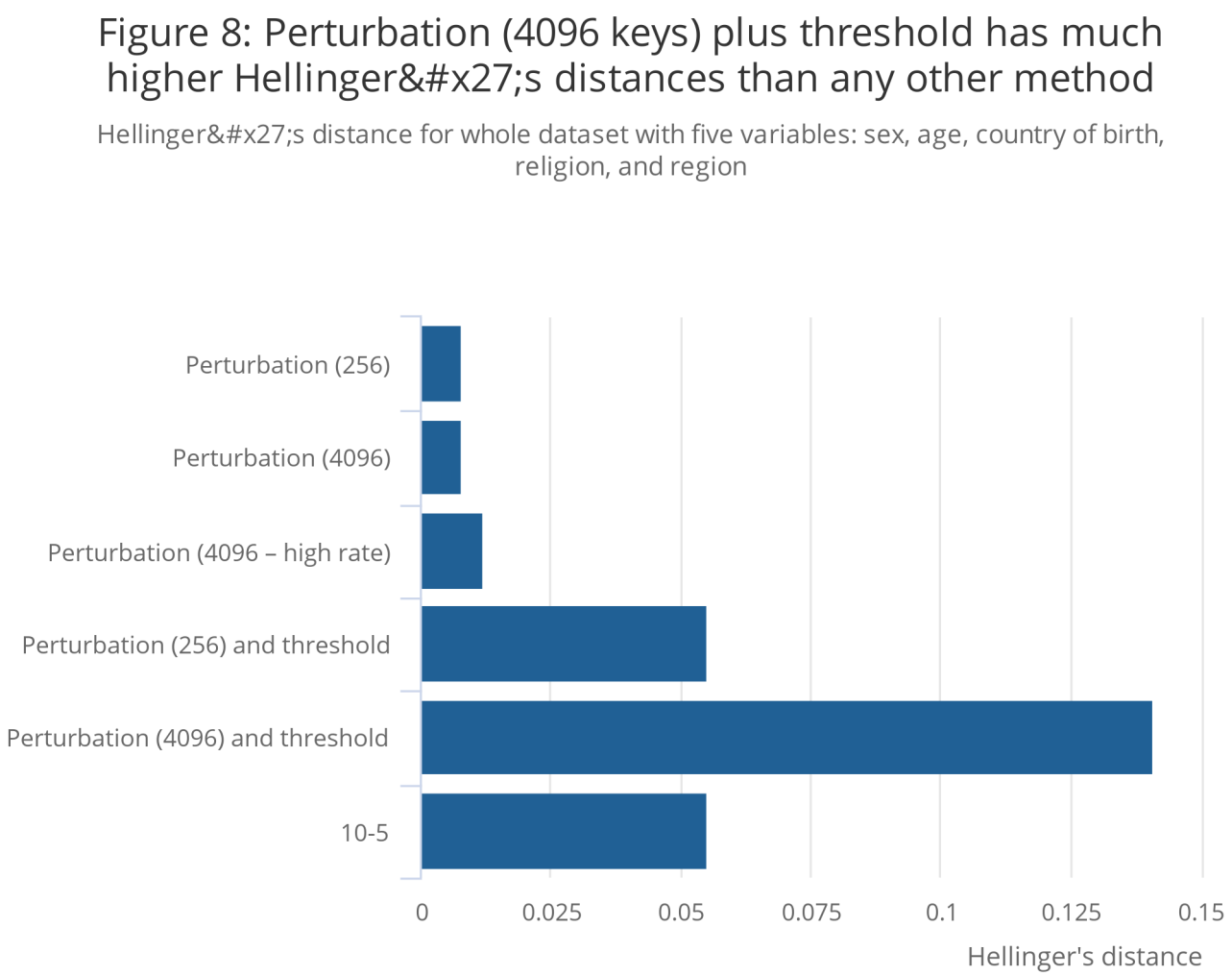
Robust eigenvalues for whole dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Figure 8: Perturbation (4096 keys) plus threshold has much higher Hellinger's distances than any other method

Hellinger's distance for whole dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Perturbation (4,096 keys) and threshold has greatly higher Hellinger's distance but lower robust eigenvalues than the other methods using a threshold.

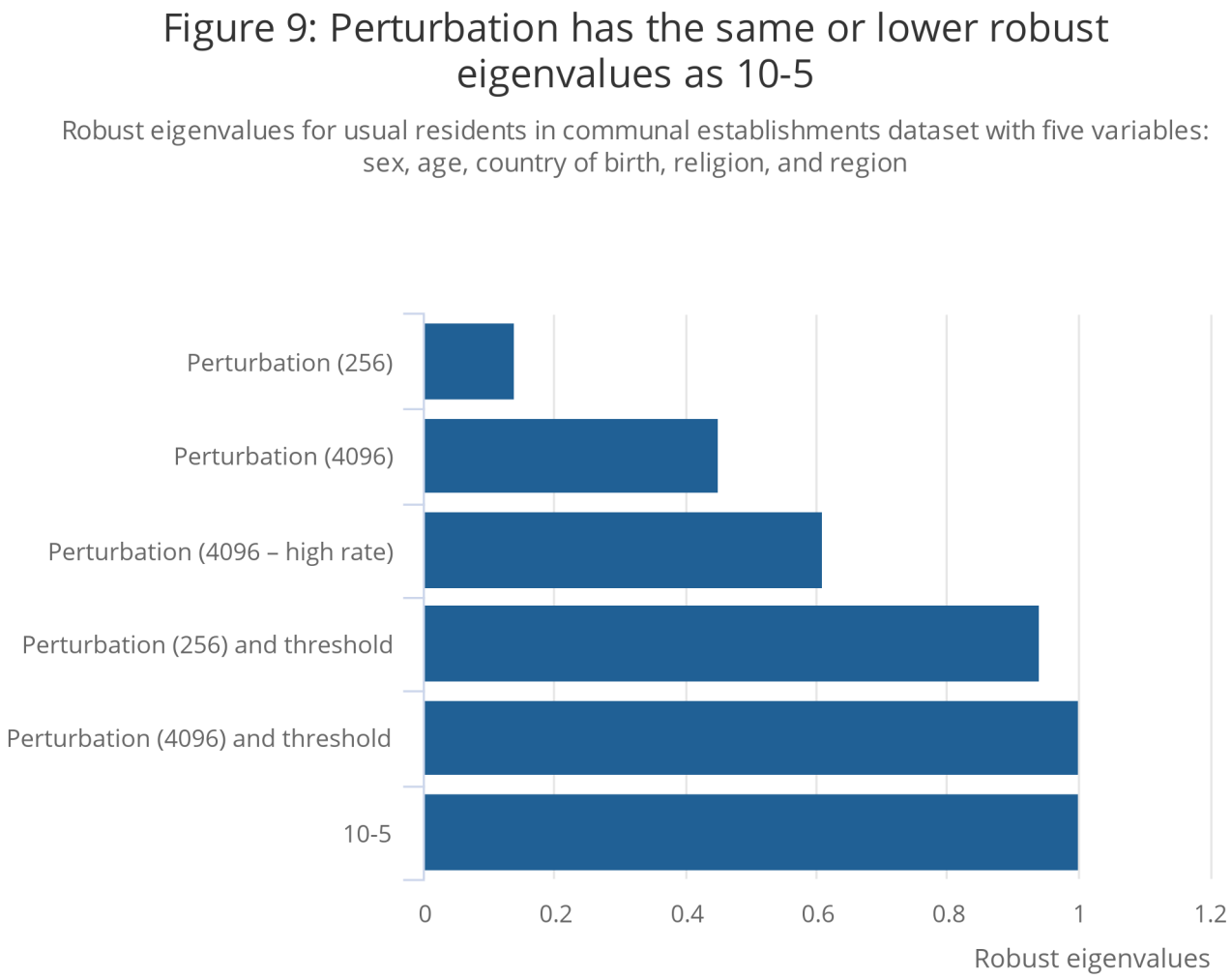
The 10-5 rule and 256-key perturbation and threshold resulted in similar values of Hellinger's distance, both near 0.055. This suggests that most of the changes were to small counts, which are affected by the threshold of 10 in both methods. Employing the 10-5 rule resulted in higher robust eigenvalues, indicating a greater level of information loss compared with perturbation.

When comparing methods without threshold, perturbation with 256 cell keys had similar values of Hellinger's distance and robust eigenvalues to perturbation with 4,096 cell keys, with high-rate 4,096 key perturbation a little higher on both measures.

To analyse how these measures compare when the dataset is very sparse, the same computations were performed on the communal establishment only dataset.

Figure 9: Perturbation has the same or lower robust eigenvalues as 10-5

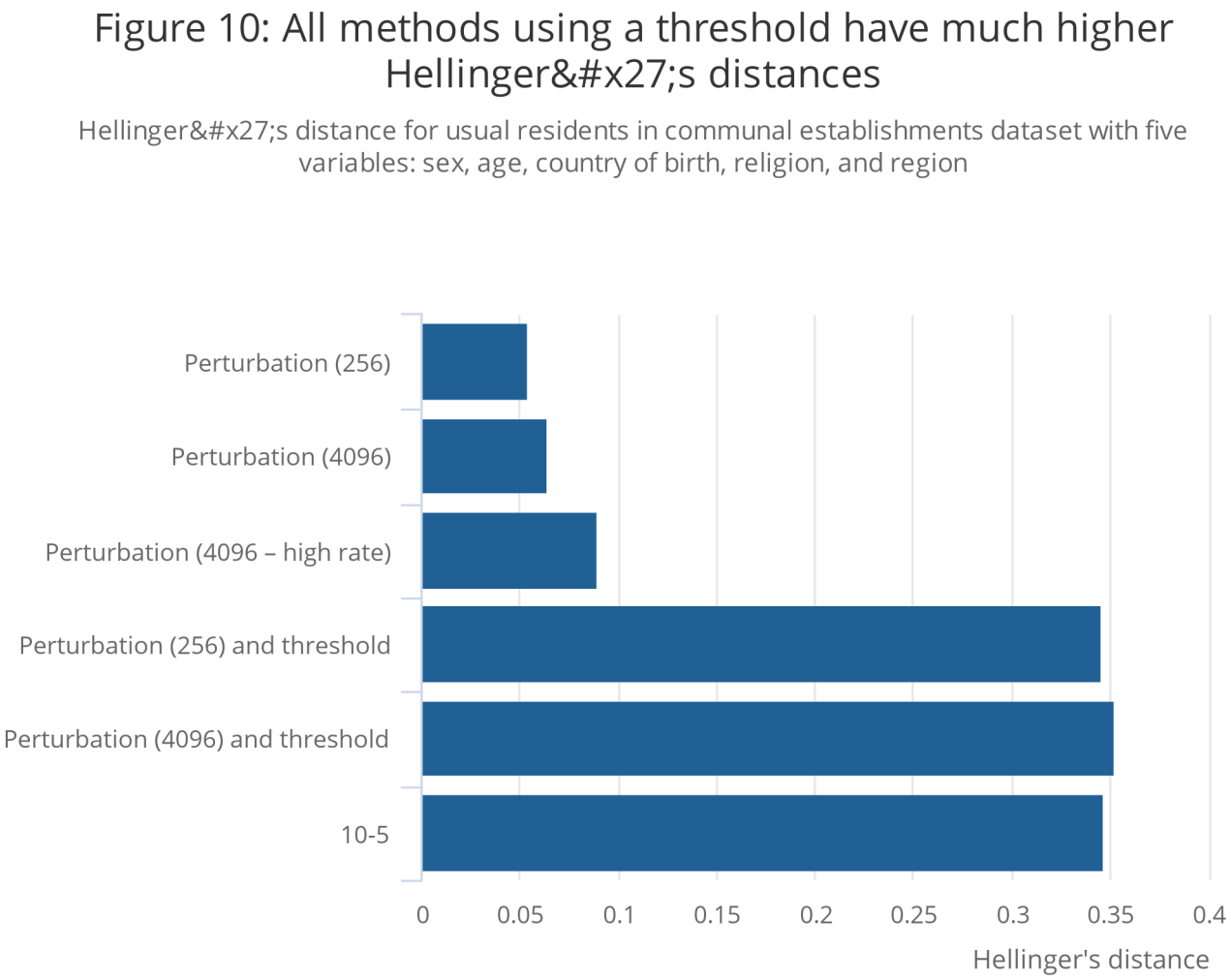
Robust eigenvalues for usual residents in communal establishments dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Figure 10: All methods using a threshold have much higher Hellinger's distances

Hellinger's distance for usual residents in communal establishments dataset with five variables: sex, age, country of birth, religion, and region



Source: 2011 Census teaching file

Similar results were observed across all methods using a threshold of 10. Perturbation with 256 keys and threshold had very slight advantages, measured by robust eigenvalues, over 10-5 and 4,096 keys perturbation with threshold.

Among methods where no threshold was applied, perturbation with 256 keys had the lowest impact on utility by both robust eigenvalues and Hellinger's distance, and high-rate perturbation with 4,096 keys as might be expected, had more impact.

7 . Conclusion

This analysis found that the use of perturbation without threshold resulted in much less noise compared with the 10-5 rule. Average noise, total noise and the percentage of overall counts changed were lower after the use of perturbation (with and without threshold) compared with the 10-5 rule. In particular, the total noise after using the 10-5 rule on the full dataset was approximately ten times higher than perturbation without threshold. Therefore, we recommend the use of perturbation when accuracy is a priority.

Applying a threshold greatly reduces the perceived disclosure risk, as no small counts would remain in the outputs. This method is also simpler to apply, but as we have explored, 10-5 incurs much more utility loss than perturbation. Still, for some analysts, rounding plus threshold may still be preferable where there are additional concerns. Context should always be considered in choice of disclosure control methods.

8 . Cite this methodology

Office for National Statistics (ONS), released 3 May 2024, ONS website, methodology, Comparison of post-tabular statistical disclosure control methods