Protecting personal data in Census 2021 results

How and why the Office for National Statistics (ONS) carried out statistical disclosure control on Census 2021 data.

Contact:
Census customer services
census.customerservices@ons.gov.uk
+44 1392 444972

Release date:
9 March 2023

Next release:
To be announced

Table of contents

1 . Main points

- By law, the Office for National Statistics (ONS) must protect the confidentiality of respondents to Census 2021.
- We protected the confidentiality of individuals' data in three ways: swapping records between areas, applying a cell key method to each table, and applying disclosure rules in deciding which tables could be published.
- These disclosure control measures have been carefully designed to protect confidentiality without distorting the statistics unduly.
- We used targeted record swapping to protect individuals and households with unusual and identifying characteristics, swapping with others of similar characteristics in nearby areas; the geographies were changed for between 7% and 10% of households, and for between 2% and 5% of individuals in communal establishments.
- We used a cell key method to protect against disclosure by differencing by adding "noise" to every dataset, known as data perturbation; a typical dataset would have around 14% of cell counts perturbed by a small amount, and small counts were more likely to have been perturbed than large counts.
- We used disclosure rules to prevent very sparse datasets in the table builder where identification and disclosure would have an increased risk.

2 . Why we use statistical disclosure control

Statistical disclosure control (SDC) covers a range of methods to protect individuals, households, businesses and their attributes, or characteristics, from identification in published datasets and microdata. Methods may be applied to the microdata (pre-tabular) or the output tables (post-tabular) before release.

The Office for National Statistics (ONS) has legal obligations under the Statistics and Registration Service Act (SRSA, 2007) Section 39 and the Data Protection Act (2018) that require the ONS not to reveal the identity or private information about an individual or organisation. The General Data Protection Regulation (GDPR) that came into force in the UK on 25 May 2018 reinforced our obligations, both in data release and data handling.

More generally, we have a pledge to respondents that the information will only be used for statistical purposes, so we must look after and protect the information that is provided to us. Moreover, a breach of disclosure could lead to criminal proceedings against an individual who has released or authorised the 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:

- is specified in the information
- can be deduced from the information
- can be deduced from the information taken together with any other published information

3 . Improvements from the 2011 Census

In the 2011 Census, the Office for National Statistics (ONS) used targeted record swapping to protect the confidentiality of individual responses and released a set of static tables. Generally, users were positive about the releases in the 2011 Census. However, concerns were raised around three aspects of dissemination: accessibility, flexibility, and timeliness. We looked to build on what worked in 2011 and address what worked less well.

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

- accessibility – this meant allowing a user to find the tables that they required; the level of detail that a user could be allowed would be subject to the assessment of disclosure risk that the combination of variables, classifications and geography would generate
- flexibility - users reported a desire to create their own outputs and frustration with decisions taken on the level of detail made available in static tables
- timeliness - users expressed disappointment that no substantial improvement had been made in 2011 compared with the release of the 2001 Census outputs

The challenge 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, we carried out work to evaluate the most appropriate combination of pre- and post-tabular methods for disclosure control. The favoured method was to consider a combination of targeted record swapping along with a post-tabular cell key method. The latter allowed us to consider the availability of an online table builder, allowing a user to find the tables that they required. The level of detail that a user could be allowed would be subject to the assessment of disclosure risk that the combination of variables, classifications and geography would generate.

In previous censuses, the policy has always been to assess whether the release of datasets is acceptable for all areas, and so every dataset that was passed was available for every area. That meant that datasets that might have been acceptable for some areas were not released because the corresponding dataset was not acceptable for other areas. This was particularly the case for some datasets with ethnic group or country of birth, where minority population groups might be geographically clustered.

Our aim for 2021 was to make datasets available for those areas where the disclosure risk would be sufficiently low, rather than reject for all areas because some might incur higher risk. We refer to the two approaches as the "blanket" approach, where datasets are produced for all areas, and the "patchwork" approach, where datasets are produced for the subset of areas where the risk is sufficiently low.

4 . Methods used

There are three methods used to protect personal data in the Census 2021 results. Each offers a complementary form of protection.

To protect individuals and households with unique or unusual characteristics, we use targeted record swapping. This prevents easy spontaneous recognition of individuals and households within datasets.

The cell key method offers protection against disclosure by differencing, where two or more slightly different datasets could be compared to expose an individual respondent, and in instances where a few datasets can be constructed and could otherwise be linked together to reconstruct records from the microdata.

The facility to "build your own" datasets could allow someone to produce an extremely large bank of datasets. The combination of these could allow identification of individuals and disclosure of information, notwithstanding the protection of "risky records" (record swapping) and the protection against disclosure by differencing (cell key perturbation). Hence, there are some disclosure rules to limit the detail.

Targeted record swapping

This is a pre-tabular method where every individual and household is assessed for uniqueness or rarity over several characteristics. Households and their individuals that are unique or rare on one or more of those characteristics are highlighted as "risky records", and all these households would be swapped.

Similar households that match on some basic characteristics are sought from other areas to be used as "swaps", to preserve data quality. These characteristics included household size, so that the numbers of individuals and numbers of households in each area are preserved.

Depending on availability of good matches, the numbers of different types of households are also preserved as much as possible.

Households are swapped within local authority districts (LADs) or, in rare cases of households with very unusual characteristics, with matches in nearby authorities. To reduce the swap rate and maintain better data utility, other risky records are prioritised for use as "matches" to be swapped with another risky household. However, where it is not possible to find a good match some "non-risky" records are used as matches, meaning that every household has a chance of being swapped.

Record swapping is also used for communal establishment data. In this case, individuals are swapped between communal establishments in different but nearby areas. The matching criteria are similar but with additions tailored to their position and the type of establishment they are in.

Cell key perturbation

Data perturbation is a technique that adds "noise" to datasets to allow individual record confidentiality. This technique allows users to find summary information about the data while reducing the risk of a security breach.

The post-tabular method is based on an algorithm that applies a pre-defined level of perturbation to cells in each dataset. The same perturbation is applied to every instance of that cell.

Firstly, a record key, which is a random number within a pre-defined range, is applied to every record in the microdata. This is done once and once only, so an individual's record key never changes.

When aggregate datasets are constructed, each cell is a count of the 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, often described 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 datasets, both instances will have the same cell value and cell key, and so receive the same perturbation. This also ensures that repeated requests of the same dataset will have the same perturbation applied consistently.

There is also some perturbation of cells where the counts are 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 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 datasets in the same way that the cell key method ensures consistency when the same cell appears in different datasets. As part of the zero perturbation, zero cells are chosen to be perturbed by, say, positive one or positive two. The same number of small cells are chosen based on category keys to be perturbed by negative one or negative two. The zero-perturbation method does not lead to any increase or decrease in overall population totals.

Note that the choice of which zeros to perturb is also based on whether the combination has appeared at a higher geography, to avoid perturbation of structural zeros. Structural zeros are cases where it would be impossible for a respondent to appear, such as the combination of "aged 0 to 4 years" and "economically active".

Disclosure rules

The disclosure checks are the rules by which decisions can be made as to whether to allow the release of outputs pertaining to specific combinations of variables. We can provide information on the rules used and some of the parameters behind them. The rationale for providing or not providing these is described within the UK Statistics Authority's Transparency of SDC methods and parameters (PDF, 277KB) report. The rules were tested rigorously against sample and exemplar datasets as an assessment of likely disclosure risk.

Small counts

In both the 2011 Census and Census 2021, it was agreed that small counts (zero, one, and two) could be included in publicly released outputs if there was sufficient uncertainty as to whether the small cell count was a true value, and that this uncertainty had been systematically created. The disclosure control methods have created that uncertainty to allow the counts to be provided. Where there is a cell count of one in a published output, there is a high chance that it could be a swapped record, a perturbed count or an imputed record, or a combination of these.

5 . The statistical disclosure control methods in practice

The data were processed in 101 delivery groups (DGs). Each DG is a set of one or more contiguous unitary authority and local authority districts (LADs), and record swapping can only take place within each DG. This means that records cannot be swapped extremely long distances and will always, at the very least, be within a region. Almost all are within the district.

The geographies were changed for between 7% and 10% of households. Between 2% and 5% of individuals were swapped between communal establishments. Fewer than 1 in every 100 swaps was between districts. These were within our tolerances; more than these levels would have damaged utility of the data unduly. We cannot provide more precise information on:

- criteria for what constitute risky records, only that they are unique or very unusual on one or more selected characteristics at a low geography
- how we matched to find swaps

This is because this information might help intruders to unpick the protection applied.

The cell key perturbation method adds controlled noise to each cell to allow more detail than previous censuses, and to allow bespoke combinations of variables, as well as those in the static datasets. The ptable was designed to add the minimum amount of noise needed for that protection. This is explained within the UK Statistics Authority's Statistical Disclosure Control (SDC) for 2021 UK Census (docx, 257KB) report. A typical dataset would have around 14% of cell counts perturbed by a small amount. Small counts were more likely to have been perturbed so datasets with large counts receive less noise than those with many small counts.

Within the "build your own" facility, the disclosure rules used are:

- the marginal minimum - where a row or column has a small total, the dataset can be susceptible to an attribute disclosure, or to help an intruder build up an individual record, if that total appears in other datasets; in deciding the value of the minimum, we take record swapping and perturbation into consideration and the likelihood of whether the records at risk are real and in the correct geographic area

- marginal dominance - a variable in a dataset should not have nearly all respondents in the same category, and there should be at least 20 respondents not in the most common category

- zeros - data should not contain an excessive proportion of empty cells; at least 40% of the dataset should be non-zero cells

- attribute disclosures - there should not be an excessive number of apparent attribute disclosures in a dataset

- sparsity (ones and zeroes) - a dataset should not contain an excessive proportion of empty cells and ones; a dataset of ones and zeroes will not only likely be risky, but it also gives a perception of risk, and if less than 50% of the dataset is non-zero cells, at least 50% of the non-zero cells should be larger than 1

- maximum number of cells - there should be an average of at least one case per cell in the dataset

- maximum number of variables - up to four variables can be selected at Output Area (OA) or Lower layer Super Output Area (LSOA) level ("below" Middle layer Super Output Area (MSOA)), and up to five variables are allowed at MSOA level and above

These disclosure checks are automated rule-based checks run by the system, which decide if there is a low enough disclosure risk to allow the release of a dataset. The rules allow release for those areas where the risk is sufficiently low, while stopping release in areas where the risk is higher. This "patchwork" approach allows more to be released than the previous "blanket" approach, which would have blocked the dataset for all areas if some areas were "too risky".

6 . Guidance when building datasets

Because we matched wherever possible on household size, the number of individuals and households in each area will be unaffected by record swapping, except for a very few cases where there was a large household that had no possible match on household size. In those instances, as close a match as possible on household size was taken.

When datasets are created, users should note that small amounts of "noise" are added to cell counts. The cell key method is intended as a "light touch" to reduce the impact of differences between totals. Where a cell containing the same records appears in the same or another dataset elsewhere, the perturbation is consistent. To allow small cells to appear in census datasets, zero-value cells are also perturbed using a similar method, which will result in a consistent perturbation of zeros.

The cell key method is to protect against using differences between similar datasets to create disclosures for small areas or specific sub-populations, known as "disclosure by differencing". The noise can be positive or negative and, across a dataset, should approximately balance out. However, the randomness may mean small changes to totals. Where two or more different datasets are constructed, the totals of all cells may in turn be different. This is because of the datasets being constructed from different cells that could be perturbed in different ways.

It is recommended, where possible, to construct the cells that you require, rather than adding up cells from a different dataset. Population totals are available in the "build your own" facility in a later release but, prior to that release, if you do need to construct totals, we recommend using totals summed from the fewest possible cells to minimise the effect of perturbation. Overall, the differences should be small and so should not change the conclusions of any analysis or research.

The disclosure rules are present to avoid users creating very sparse datasets in the "build your own" facility that might make identification either straightforward or at least increase the risk of linking between datasets to allow a disclosure. Where a user is unable to access a dataset for some geographic areas because of the disclosure rules in the table builder, we will not provide details of which rule or rules have caused the areas to fail. However, we recommend that the user considers either reducing the detail for one or more of the variables or using a higher geographic level.

The rules in the table builder are necessarily more cautious since they are automated, and some datasets (areas) that do not pass may be available in the Ready-made datasets. We cannot assess every combination of variables and classifications manually. However, the limited number of Ready-made datasets has received a closer inspection that allow us to consider context and some other aspects that might offer protection; that is, a "closer inspection" that cannot be automated.

7 . Related links

Statistical Disclosure Control (SDC) for 2021 UK Census (docx, 257KB)
Report | EAP125 | Released October 2020
This UK Statistics Authority report presents the SDC methods for approval at UK Census Committee.

Transparency of SDC methods and parameters (PDF, 277KB)
Report | EAP168 | Released September 2021
This UK Statistics Authority report is a post-meeting External Assurance Panel (EAP) publication for SDC for Census, August 2021. It summarises the SDC methods and discusses transparencies of the methods and parameters.

8 . Cite this methodology

Office for National Statistics (ONS), released 9 March 2023, ONS website, methodology, Protecting personal data in Census 2021 results