Census 2021

Statistical Disclosure Control Methodology

Guidance note

Version 3: June 2023

Northern Ireland
Statistics and Research Agency

Gníomhaireacht Thuaisceart Éireann
um Staitisticí agus Taighde

1. Introduction

Statistical Disclosure Control (SDC) refers to a range of methods that aim to protect individuals, households, businesses, and their attributes from being identified in published data and microdata. Methods may be applied to the microdata (pre-tabular) or the output tables (post-tabular) prior to release.

Northern Ireland Statistics and Research Agency (NISRA) have legal obligations under the Statistics and Registration Service Act (SRSA, 2007) Section 39 and the Data Protection Act (2018) requiring that NISRA do not 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. See NISRA privacy notice for further information.

2. Statistical disclosure control methods

For Census 2021, NISRA have applied three strategies to ensure individuals are protected from identification while minimising the impact on the quality of results:

- Targeted Record Swapping

- Cell Key Perturbation

- Disclosure rules within the Flexible Table Builder system

2.1. Targeted Record Swapping (TRS)

Targeted Record Swapping is a pre-tabular method of disclosure control used to protect individuals and households with unique or unusual characteristics. This method can achieve a high degree of confidentiality while maintaining the integrity of the data.

2.1.1. Census 2021 application of Targeted Record Swapping in households

All households and the individuals within them were assessed for risk based on a small number of characteristics that are considered unique or rare, and households that were most at risk of identification were swapped with similar households from other geographical areas.

Similar households were matched on basic characteristics so data quality could be preserved (i.e., household size so the number of people in households in each area is

unspoiled). Any individual or household could be selected for swapping; however, those deemed riskier were more likely to be swapped. Households were swapped within Local Government Districts (LGDs) or, in rare cases of households with very unusual characteristics, with matches in nearby LGDs.

2.1.2. Census 2021 application of targeted record swapping in communal establishments

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

2.2. Cell Key Perturbation (CKP)

Cell Key Perturbation is a post-tabular method of disclosure control that involves adding "noise" to some cells in a table. This adds uncertainty to the table with the goal of protecting against disclosure. The method uses an algorithm to apply a pre-defined level of perturbation to cells in each table, with the same perturbation being applied to every instance of that cell. Small counts are more likely to be perturbed, so tables with large counts have less noise added than those with a lot of small counts. The noise can be positive or negative and should approximately balance out across a table.

The cell key method provides protection against using differences between similar tables to create disclosures for small areas or specific sub-populations, known as "disclosure by differencing". The method also allows more detail than previous censuses and allows bespoke combinations of variables.

Cell key perturbation is applied to all published Census 2021 tabular outputs. In addition, cell key perturbation is embedded within the newly developed Flexible Table Builder so that a table created in the builder will give identical results to the same pre-published table.

2.2.1. Census 2021 application of Cell Key Perturbation

Firstly, a random number within a pre-defined range (known as the record key) was assigned to every record in the microdata. The record key was applied only once, and therefore never changes for an individual.

A lookup table (called the ptable) was then used to determine the amount of perturbation that should be applied to cells in the published data tables. When a table was constructed, a

cell key was calculated for each cell from the sum of the record keys within. The cell-key value was then sought in the ptable, and the required perturbation added. Where the same cells appeared in a different table, both instances had the same cell key and so have the same perturbation.

There was also some perturbation of cells where the counts were zero. A random number was assigned to each category of each variable and used to produce a random and uniformly distributed category key. The category key was used to make a random selection of cells to perturb. The perturbing of zero counts was done in a consistent manner, while ensuring that structural zeros remain unaffected. The method to perturb zeros does not lead to any increase or decrease in overall population totals; if zero cells are chosen to be perturbed by, say, positive one or positive two, the same number of cells are chosen to be perturbed by negative one or negative two.

2.3. Disclosure rules within the Flexible Table Builder system

To prevent large and sparse tables being accessed from the build-a-custom-table facility in the Flexible Table Builder, disclosure rules have been set within the system. These disclosure rules ensure sensitive outputs are protected from risk of disclosure and identification while providing users with the flexibility and rapid access to tabular outputs they desire.

When a table is created in the builder, the system checks against the pre-defined disclosure rules to determine whether the outputs are too disclosive to release. The rules are checked for each specific combination of variables and geographic areas the user attempts to build for, withholding data for those areas that do not meet the requirements. This "patchwork" approach allows greater accessibility than the "blanket" approach used in previous censuses, where data for all areas would have been blocked if any areas were deemed too risky.

2.3.1. Census 2021 application of disclosure rules

The disclosure rules used within the Flexible Table Builder include:

- the marginal minimum - where a row or column has a small total, the table can be susceptible to an attribute disclosure, or to help an intruder build up an individual record, if that total appears in other tables; in deciding the value of the minimum,

record swapping and perturbation was taken into consideration and the likelihood of whether the records at risk were real and in the correct geographic area.

- marginal dominance - a variable in a table 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 table should be non-zero cells.

- sparsity (ones and zeroes) - a table should not contain an excessive proportion of empty cells and ones; a table 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 table 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 table.

- maximum number of variables - up to four variables can be selected at Data Zone (DZ), Super Data Zone (SDZ) level, District Electoral Area (DEA) and Settlement, and up to five variables are allowed at Local Government District level and above.

3. Small counts

In both the 2011 Census and Census 2021, small counts (zero, one, and two) can be included in publicly released outputs if there is sufficient uncertainty as to whether the small cell count is a true value, and that this uncertainty has been systematically created. The disclosure control methods have created this 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.

4. Guidance when building tables in the Flexible Table Builder

As a result of the statistical disclosure controls being applied to Census 2021 data prior to users gaining access, the following guidance should be noted when producing tables in the Flexible Table Builder:

- When tables are created, 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 table elsewhere, the perturbation is consistent. Zero-value cells are perturbed using a similar method, which will result in a consistent perturbation of zeros.

- The noise resulting from cell key perturbation can be positive or negative and, across a table, should approximately balance out. However, the randomness may mean small changes to totals. Where two or more different tables are constructed, the totals of all cells may in turn be different. This is because of the tables being constructed from different cells that could be perturbed in different ways.

- It is recommended, where possible, to construct the table that you require directly from the Flexible Table Builder system, rather than adding up cells from a different table. Otherwise, using totals summed from the fewest possible cells to minimise the effect of perturbation is recommended. Overall, the differences should be small and so should not change the conclusions of any analysis or research.

- Where a user is unable to access a table for some geographic areas because of the disclosure rules in the Flexible Table Builder, details will not be provided of which rule or rules have caused the areas to fail. However, users are advised to consider either reducing the detail for one or more of the variables or using a higher geographic level.

These disclosure control measures have been designed to protect confidentiality without distorting the statistics unduly. In consideration of the statistical disclosure control methodology, Census Office aimed to strike a reasonable balance in providing utility, accessibility and flexibility to users while also fulfilling legal obligations to protect against disclosure risk.