

Disclosure Control for Census 2021: Making Census Safe!

Census 2021

Statistical Disclosure Control

15 May 2023

Thank you for your interest

There will be time for questions at the end: please post in the chat as we go

What does safe mean?

It means you can't identify people in the data, or learn things about them you didn't already know

SDC Methods for Create a custom Dataset 2021

A short account of the process

Statistical Disclosure Control (SDC) methods

- **Targeted Record Swapping** – identifying people and Households that stand out in the data, and swapping them with a similar record in the local area
- **Cell Key Perturbation** - this adds noise to the figures, making slight changes to cell counts
- **SDC rules** – this stops tables with many zeros/ low counts from being released

Targeted Swapping

- Any record could be swapped, but rarer ones are much more likely to be
- Swapping stays in the local area – we swap between OAs, or between MSOAs in the same Local Authority
- Increases uncertainty over any identification

Everyone who answers the Census is still present, just not necessarily in the same place.

Cell Key Perturbation

- This makes slight changes to the figures, adding or subtracting
- This adds uncertainty about any identification from a small count (including zeros)
- It makes it harder to add up totals to work out missing data (differencing)

The purpose is to prevent disclosure without making the data less useful

SDC rules

- If a table has too few records in it, it will be suppressed
- There are specific rules to protect communal establishments, households, and individuals
- These rules are built into [Create A Custom Dataset](#)

The rules are designed to prevent

- Attribute disclosure
- Too many low cell counts
- Dominance
- Building up a picture from repeat queries

Why don't totals add?

Hobbiton	0-15	16-24	25-44	45-115	Total
All	9	11	26	18	64

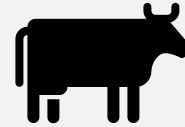
Hobbiton	
male	35
female	28
Total	63

If you get the same items in a cell, the cell value is the same..

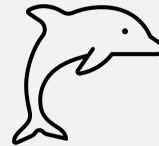
But in different tables, you get different totals

Things the rules protect against:

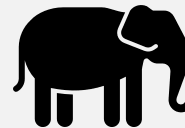
Industry	0-15	16-24	25-44	45-115
Milk production	0	0	0	2
Bakers' Shop Workers	0	0	5	10
Swim Teacher	0	0	0	6
Farmer	0	3	0	12



How old is Ermintrude?



How old is Danny?

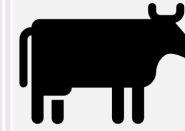


Gerty is 26, what is her job?

The table builder may suppress tables like this

But the table builder might give more detail:

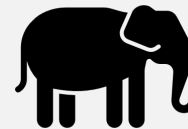
Industry	0-15	16-24	25-44	45-64	65-115
Milk production	0	0	0	2	1
Bread baking	0	0	5	5	5
Swim Teacher	0	0	0	4	2
Farmer	0	3	0	6	6
Bread Sales	0	2	2	2	1



How old is Ermintrude?



How old is Danny?



Gerty is 26, what is her job?

Intruder Test

- Friendly intruders tried to identify people
- Most were unsuccessful
- Confirmed there were risks with detailed classifications at low geography



Table Outputs

A matter of judgement

15 May 2023

Considerations

- Sensitivity
- How many low counts
- Purpose of the data
- How well other considerations protect it (imputation, swapping, perturbation)
- Level of geography versus level of detail

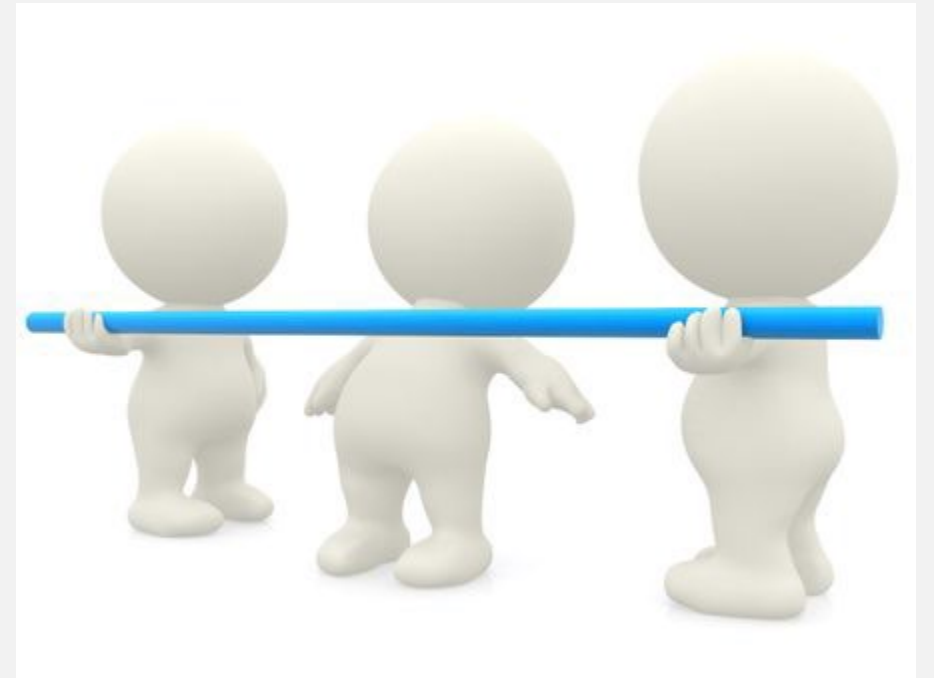


Microdata

Record level data available through varying channels

Public File

- 1% sample
- 20 variables
- Region Geography
- Low Detail



Risk level: LOW

Safeguarded HH 1%

- New for 2021!
- Geography : Region
- 60 (ish) variables

Risk Level:
Medium

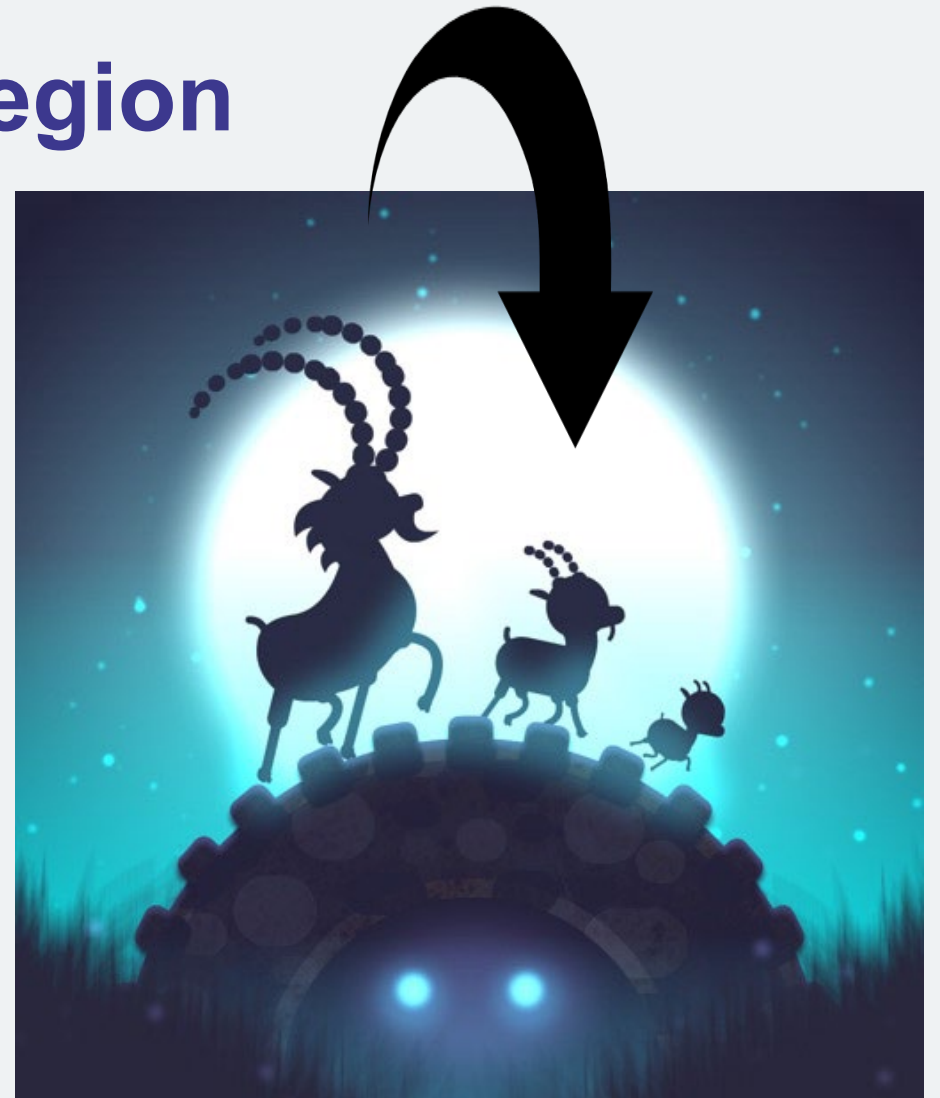


This Photo by Unknown Author is licensed under [CC BY-NC-ND](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Safeguarded Individual 5% Region

- Detailed variables
- 90 (ish) variables
- For a guide see [2011 codebook](#)

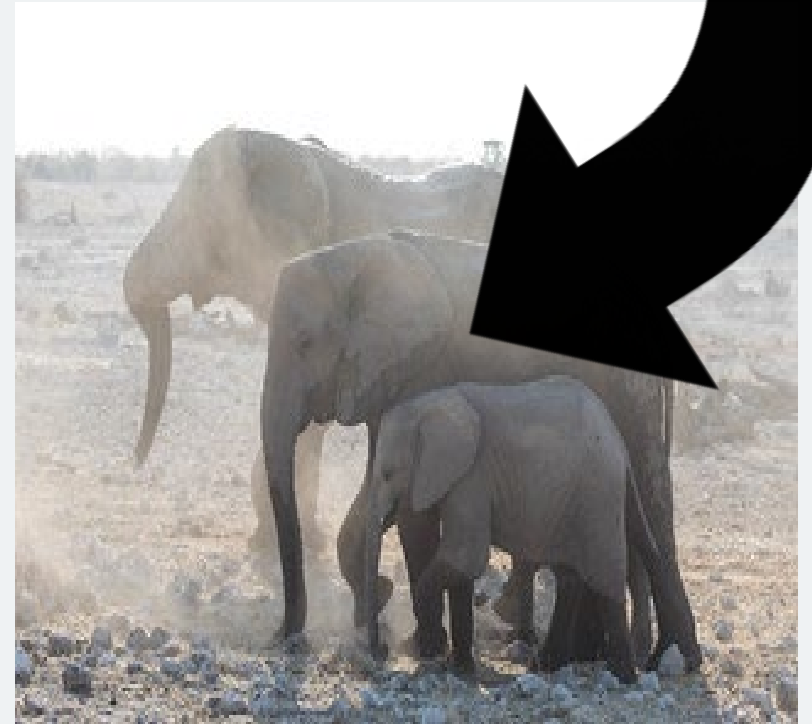
Risk Level:
Medium



Safeguarded Individual 5% Grouped Local Authority

- Grouped Local authority: every Local Authority under 120k head of population grouped up
- 90 (ish) variables
- Low Detail
- Please see the [2011 Codebook](#) for a guide

Risk level: Medium



Secure Individual 10% Local Authority

- Available through Integrated Data Service
- Possible to cross reference with a wide range of variables – see the [Census 2011](#) datasets for a guide
- 120+ Variables

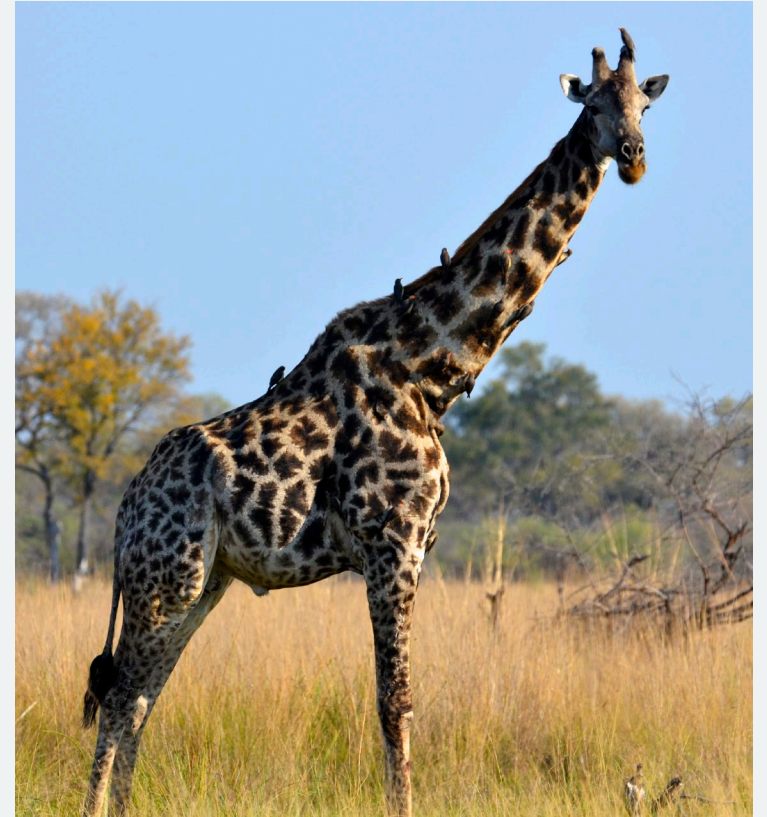


Risk Level:
HIGH

Secure Household 10% Local Authority

- Available through IDS
- Possible to cross reference with wide range of other variables
- Data linkages possible with wide range of other resources
- Near 200 Variables
- Please see the [2011 datasets](#) for a guide

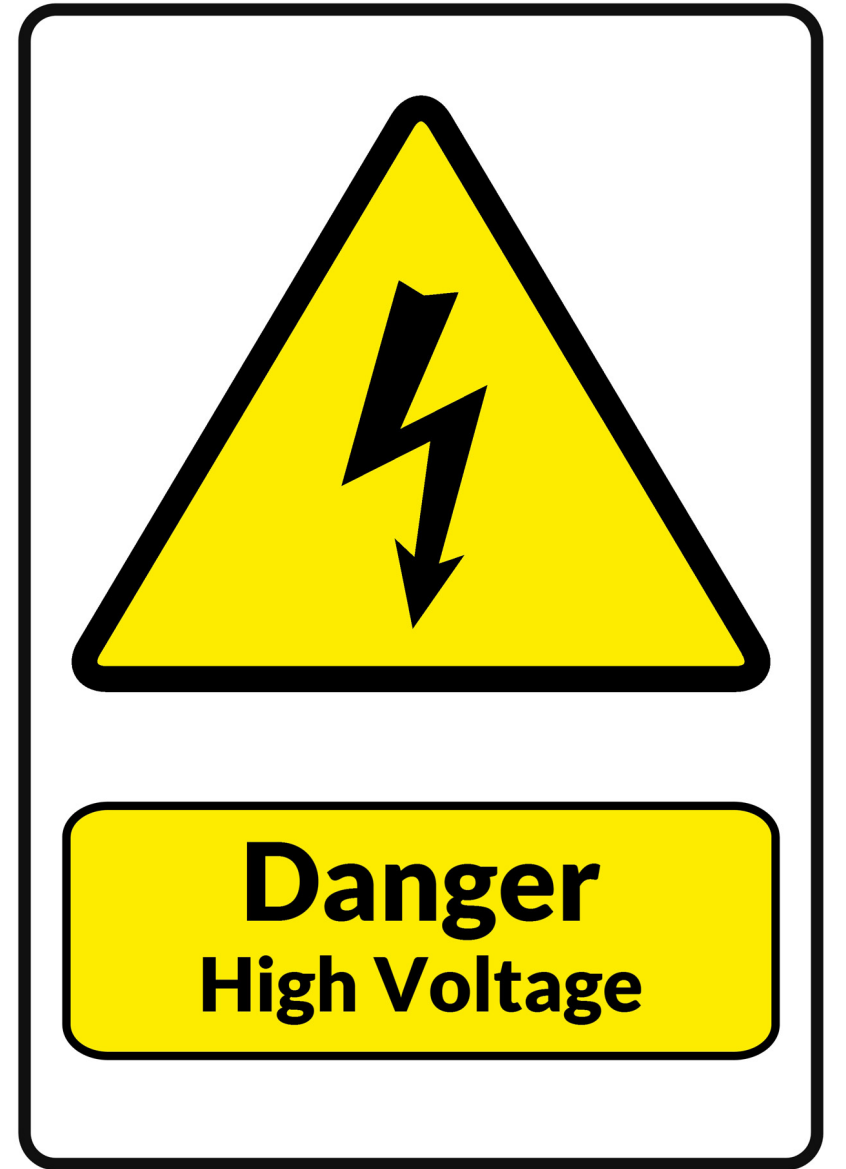
Risk Level: HIGH



Secure 100% HH & Individual

- Project must justify need for 100% data
- Great Potential for data linkages
- ‘Proof of major Policy Impact’
- See [2011 prospectus](#) for ideas

Risk Level:
Highest



Considerations

- User need
- Right detail for the right dataset
- Intruder Testing!

