Census outputs and Statistical Disclosure Control

Create a custom dataset – MARP briefing note April 2023

On Tuesday 28th March, the census 2021 Create a custom dataset tool was released, allowing users to select their own combinations of variables in datasets. Details of the disclosure control applied to these releases have been previously discussed with the Methodological Assurance Review Panel here: Transparency of SDC methods and parameters (PDF, 277KB) and the main parameters of perturbation and the disclosure checks have been published: Protecting personal data in Census 2021 results - Office for National Statistics (ons.gov.uk).

Create a custom dataset is a useful tool for creating bespoke outputs on a combination of variables. It aims to provide much more flexible outputs that can be provided rapidly using a series of automated checks.

However, census data are provided in a range of products:

- analytical articles
- topic summaries
- multivariate data
- alternative and small population data
- origin-destination (flow) data
- microdata samples
- research via secure access
- commissioned tables (and others)

Each output route serves a different purpose (more information can be found on the census products page). Custom datasets will not suit every user need and some data are better served by other output routes.

Not all variables and classifications were included in create a custom dataset. The criteria considered for what went into the website include:

- Disclosure Risk
- Utility
- User Burden
- ONS Resource
- Software Constraints
- User need / Priority

The most important criterion for inclusion in the flexible system was that the custom outputs on the data would be safe for release, given the protection provided by the census SDC methods: record swapping, cell key perturbation, and the disclosure rules.

Some variables are not well suited to the flexible system due to the small populations of some of their categories. The most common reason for a variable not being included in the system was that the system would provide very low counts that would fail the disclosure rules and provide a poor user experience. For example, providing a very detailed classification with several hundred categories would lead to low counts in some categories even at national or regional level, and would fail the disclosure checks for most requests. It was important to avoid over-promising or suggesting that an unrealistic level of detail would be available, and prevent the resulting situation where a user repeatedly requests datasets including very large classifications, but the system always fails the areas due to low counts.

On a similar note, there is considerable user demand for data on small populations, for example, the characteristics of all members of a particular religion or ethnic group. Although some data on this topic will be available, the create a custom dataset system returns data for the whole population and every category in a classification. Requests cannot be made for a single religion or ethnicity without requesting data for every religion or ethnicity in the classification, which are likely to contain small counts and cause failures of the disclosure checks. For this reason, user need for data on members of particular groups will often be better met by ‘alternative and small population’ outputs which can produce datasets containing only the population of interest. Allowing requests to be made on parts of a classification only, or on individual categories, would circumvent the disclosure rules which were designed to assess the risk of tables with complete classifications.

Each request made to the flexible system is assessed using the same disclosure rules, and each request is assessed independently. The availability of other data is not considered in decisions on which areas pass or fail the disclosure checks. Although the availability of other, similar data would inform the potential disclosure risk, implementing a system that considers the results of previous requests would be technically challenging. It was also preferred that each request was assessed based on its own merits and that selections of early users did not affect how much data was available to later users, in a ‘first come first served’ manner.