Modernising statistical data dissemination with the Office for National Statistics in the United Kingdom

Challenge

Applying statistical disclosure control techniques to **census-scale survey data** in order to produce a range of outputs that are safe to release takes a great deal of time and effort. This limits the range of outputs that can be produced and creates a time lag between the end of a survey and the publication of outputs.

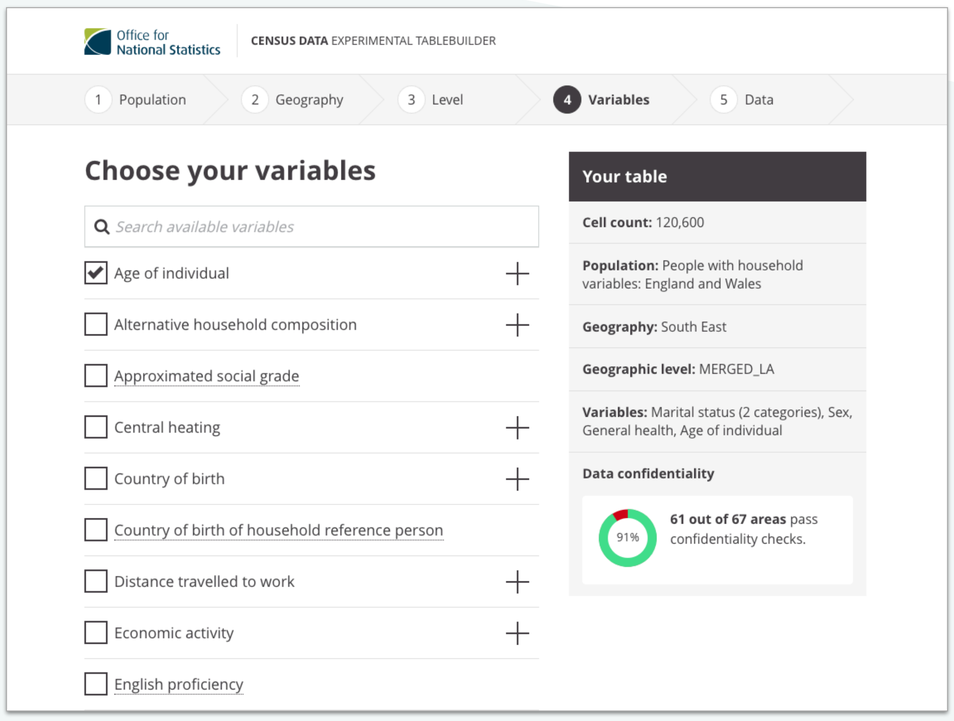


Figure 1: Screenshot of Cantabular software

Benefits

More data

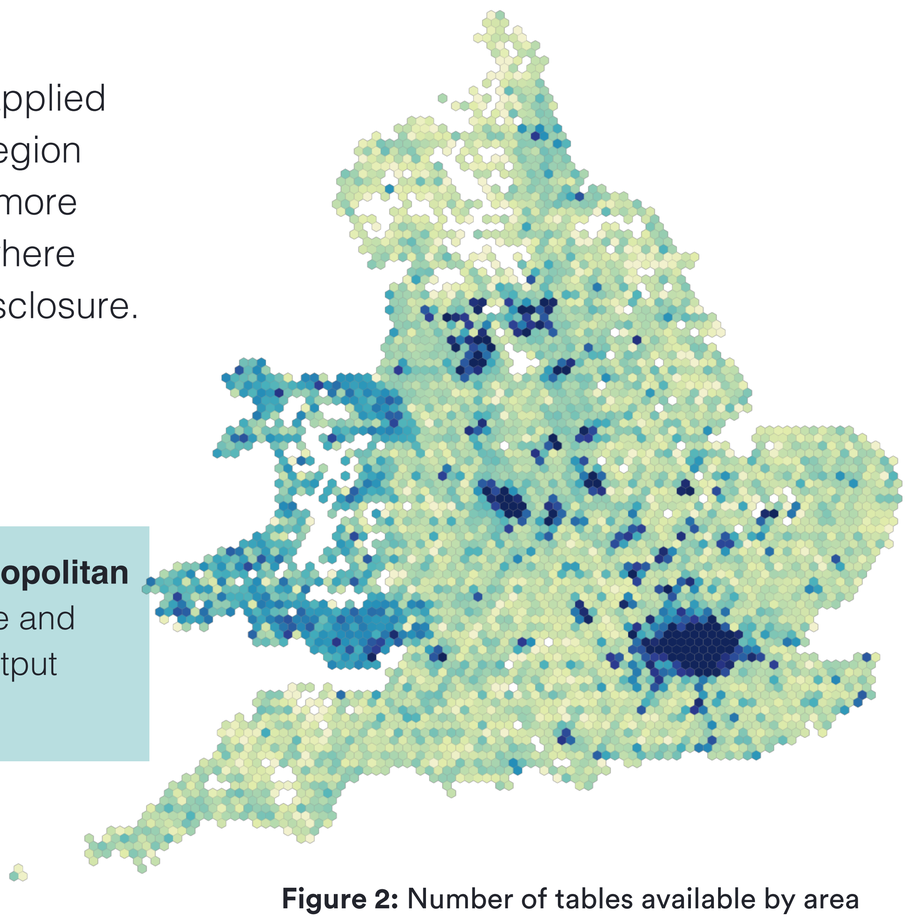
Users can precisely define their own queries instead of relying solely on pre-published outputs.

More quickly

Automatically checked outputs allow the time between collection and publication to be shortened.

For more areas

Disclosure rules are applied to each geographic region individually, allowing more data to be released where there is less risk of disclosure.



Darker, typically metropolitan areas are more diverse and have more possible output tables available.

Figure 2: Number of tables available by area

Solution

The Sensible Code Company worked with the ONS to develop dissemination software for use in the 2021 Census. The system **applies disclosure control techniques in response to a user's request in real time**. It takes in row-swapped microdata, performing aggregation, applying perturbation and checking the output against disclosure rules.

Innovative technology

Built in a memory-safe language, with both security and performance as up-front concerns. An application-specific database enables many concurrent users to query more than 60 million rows of data in real time. The system can easily be scaled up with additional servers to meet any peaks in demand that may occur.

Computing all tables

Cantabular is fast enough to compute all allowed tables (those which pass the rules). This facilitates novel user interfaces for exploring published data. It would also allow for deployment outside a secure environment.

Disclosure control approaches

The ONS use geographic row-swapping before loading data into Cantabular. Three additional methods are then used to protect privacy.

Cell key method

Pioneered by the Australian Bureau of Statistics, this method perturbs cell counts using random secrets generated when a dataset is built. This disrupts inference about contributors to a cell while ensuring a repeated query returns the same result.

Structural zeros

These are impossible combinations of categories, such as married children. Cantabular automatically infers structural zeros from the data by performing a parallel query at a coarser geographic level to avoid perturbing these cell values.

Disclosure rules

After perturbation, outputs are evaluated against a set of rules which test likely disclosiveness. Tests are run for each geographic sub-area. Data is withheld where the sub-area table fails any of the rules. Rule sensitivity can be adjusted by data controllers.