

International Journal of Population Data Science

Journal Website: www.ijpds.org



Developing a generalisable stratification approach for clerical review of linked data

Leah Maizey¹, Josie Platcha¹, Tim Gammon¹, Matt Wray¹, Gavin Thompson¹, Laszlo Antal¹, and Rosaland Archer¹

¹Office for National Statistics

Objectives

Data linkage is a vital process in the creation of many national statistics, but understanding the quality of linked data is currently highly inefficient. To find errors, data must be reviewed by humans which is costly and lengthy. Sampling is used to reduce the clerical burden. This research aims to develop a method for stratifying links to create representative samples while reducing the number reviewed. The final method will enable nuanced stratification of data for review whilst optimising resource efficiency.

The objectives are to:

1. ensure that the method is adaptable across diverse datasets,
2. achieve full automation,
3. ensure scalability to accommodate large datasets.

Approach

Our approach centres on designing an algorithm that responds to the variability in the data distribution of probabilistic scores and stratify accordingly. The intention is for the developed method to automatically adjust its parameters, such as strata threshold and numbers based on the data's characteristics.

The research involves a comparative analysis of the performance of dynamic- and percentile-based stratification against the current standard practice of static threshold stratification.

Results

Tests are ongoing to compare the above methods on a variety of metrics including homogeneity of strata, total variance, and between-strata distance. Findings will be presented at the conference.

Conclusions

We hope to design a robust, generalisable and scalable stratification method that can be integrated into a Linkage pipeline.

Implications

Implementing the method will help to improve the quality of national statistics, ensuring more accurate, reliable and timely outputs are produced in a resource efficient manner.

