International Journal of Population Data Science

Journal Website: www.ijpds.org

A method for Linking Multiple De-identified Datasets

Waugh, A^{1*}, Rowley, D¹, and Clarke, A¹

¹National Records of Scotland

Background

National Statistics Institutes have been exploring the value of using administrative data. The Administrative Data Team within the Scotland's Census 2021 Programme are exploring bringing administrative datasets together to support the census and produce alternative population estimates.

Objectives

We are developing methods to link de-identified administrative datasets, drawing on existing methods.

Methods

Our method uses hashed linking variables, derived from name, address, date of birth and gender. One linking variable is a names correction, produced by comparing names to each name in a reference set and scoring the difference. The scoring algorithm developed considers transpositions, deletions, insertions, substitutions and moves, and is sensitive to the particular letters involved.

Linking variables are combined at run time to produce thousands of matchkeys, allowing more matches to be linked deterministically using hashed data. Overall link strength scores are calculated as a combination of:

- Penalties associated with the matchkey, based on the linking variables used, and
- Similarity on dates of birth, measured at run time using weighted Bloom Filters.

We concatenate all the datasets and link the resulting dataset to itself. This allows simultaneous linking across all datasets and resolution of duplicate records within each dataset. This results in potentially complex patterns of links. By considering the records and links as a graph we allocate records to unique individuals through a vertex colouring algorithm on the complement of each component. The link strength is considered to prioritize allocation.

Findings

Clerical review on links made found that those with stronger scores were more likely to be considered a match.

Conclusions

This linking method is being used and tested further in linking admin datasets for population estimates. We also plan to use it for several linking tasks in the processing of Scotland's Census 2021.

*Corresponding Author:

Email Address: andrew.waugh@nrsotland.gov.uk (A Waugh)