

International Journal of Population Data Science

Journal Website: www.ijpds.org



Identifying Scottish siblings: A population-scale approach to link historic birth, marriage, and death certificates

Charini Nanayakkara¹, Peter Christen^{1,2}, Chris Dibben², Eilidh Garrett², Fiona Hemsley-Flint², and Lee Williamson²

¹Australian National University, Canberra, Australia

²University of Edinburgh, Edinburgh, United Kingdom

Objectives

Reconstructing populations by linking vital event records can facilitate a variety of studies including the analysis of hereditary illnesses and socioeconomic changes. We present a record linkage framework to identify siblings, which is a first step in population reconstruction, applied on a Scottish database spanning nearly 120 years.

Methods

Pairwise comparison of records to identify matches (siblings) and non-matches (non-siblings) is not scalable to large population databases containing millions of records. We therefore apply a novel blocking approach based on Locality Sensitive Hashing to reduce the comparison space, and employ multiprocessing techniques to further improve scalability. In addition to calculating attribute similarities to determine whether a record pair is a match or not, to improve linkage quality we also incorporate temporal constraints (such as siblings born three months apart being not feasible). The final linkage results are stored in a Neo4j graph database to facilitate querying and visualisation.

Results

We apply our record linkage framework on vital event records (around 14 million birth, 8 million death, and 4 million marriage certificates) from Scotland to identify records that correspond to sibling groups. We generate a similarity graph, with nodes representing records and edges corresponding to similarities, by comparing over 150 million record pairs using attributes that are expected to be similar for siblings (such as mother's name and parents' marriage place). Using graph-clustering techniques we then group records such that each cluster represents a sibling group. We independently link birth, death, and marriage certificates for sibling group generation to create complementary results, which we then use to identify high confidence links. We also employ unsupervised evaluation techniques to assess the quality of our linkage results.

Conclusion

Large-scale population record linkage is non-trivial due to quality and scalability challenges. We propose a scalable and effective population linkage framework for identifying siblings by linking and clustering vital event records. We store our linkage outcomes in a graph database to facilitate visualisation and research based on reconstructed populations.

