International Journal of Population Data Science

Journal Website: www.ijpds.org

Bias, accuracy and sample size in the systematic linking of historical records

Antonie, L¹, Inwood, K¹, Minns, C², and Summerfield, F³

¹University of Guelph

²London School of Economics

³St Francis Xavier University

Introduction

Linking distinct historical sources on an automated basis directs attention to the quality and representativeness of the linked data created by these systems. Linking with time-invariant personal characteristics arguably minimizes bias or departures from representativeness even though a wider set of features might generate more links.

Conclusion/Implications

Results suggest that neither method will be universally preferable. Rather, the choice of research question may affect the preferred balance of biases and link rate. Fortunately, the advance of computational capacity allows a researcher to select a method that generate links most appropriate for the problem at hand.

Objectives and Approach

The objective of this research is to compare, evaluate and understand bias when two linking methodologies are employed on the same data sources. Our approach to studying this problem is by comparing linked records from Canadian censuses (linking 1871 to 1881) generated by two different linking strategies. The first method is a support vector machine based classification model on time-invariant individual characteristics. Using this method a large number of multiple matches is generated, as records look similar on a small number of time-invariant individual characteristics. The second method adds a second stage of disambiguating multiple matches using family information.

Results

We compare the links produced by the two methods used in the study and we discuss the results. The comparison is in terms of number of links produced, their quality (false positive rate) and the bias of the linked data produced. A complication is that there are many dimensions of bias. Even time-invariant criteria typically generate some bias.

As expected, the two-step process produces a larger linked sample. Interestingly, it also produces a lower error rate and different patterns of bias. Both methods understate the Quebec-born, French-ethnicity, the unmarried and adolescents. Unexpectedly, the bias in favour of married people is larger using individual (first method) than family information (second method). However, family-based linking does over-represent young children.