International Journal of Population Data Science

Journal Website: www.ijpds.org

Better than blocking: building and using multiple variable pattern indexes (MVPi) to screen data sources for possible matches and reduce linkage search space

Hermansen, S¹

¹Westat

Introduction

Scaling up data linkage presents a challenging problem that has no straightforward solution. Lacking a prescribed ID in common between two data sources, the number of records to compare increases geometrically with data volume. Data linkers have for the main part resorted to “blocking” on demographic or other identifying variables.

Conclusion/Implications

The screening phase of a large-scale linkage project reduces linkage search space to the pairs of records more likely to be true matches, but it may also lead to selectivity bias and underestimates of the sensitivity (recall) of data linkage methods. Efficient MVPi screening allows fuller use of identifying information.

Objectives and Approach

Among the more efficient of better blocking methods, carefully constructed multiple variable pattern indexes (MVPi) offer a robust and efficient method for reduction of linkage search spaces in Big Data. This realistic, large-scale demonstration of MVPi combines 30,156 SSA Death Master File (DMF) and NDI matches on SSN with equal dates of death (true matches) and 16,332 DMF records with different or missing SSN, and links the total of 46,448 records on names, date of birth, and postal code (ignoring SSN) to >94MM DMF records. The proportion of true matches not linked tests for loss of information during the blocking phase of data linkage.

Results

Blocking has an obvious cost in terms of completeness of linkage: any errors in a single blocking variable mean that blocking will miss true matches. Remedies for this problem usually add more blocking variables and stages, or make better use of information in blocking variables, requiring more time and computing resources. MVPi screening makes fuller use of information in blocking variables, but does so in this demonstration in one cohort (>30K) and DMF (>94MM) data sets. In this acid-test demonstration with few identifying variables and messy data, MVPi screening failed to link less than eight percent of the cohort records to its corresponding true match in the SSA DMF. MVPi screening reduced trillions of possible pairs requiring comparisons to a manageable 83MM.