Matchmaking gone wrong: Quantifying bias and methods using non-western data

Alexander Persaud

apersaud@richmond.edu

The construction of panel data sets using census or other similar data has enabled researchers to examine intergenerational outcomes or to quantify the long-run effects of early interventions. (See Long and Ferrie (2013), Clark (2014), (Bleakley and Ferrie (2016), and Aizer et al. (2016).) However, incorrect matching, both false matches and false non-matches, may affect the results in these and similar papers. Bailey, Henderson, and Massey (2017) show that different assumptions, methods of matching, and breaking ties lead to results of varying quality. One key limitation in the matching literature is the lack of a truth sample, i.e., data with true matches. Another limitation is the large gap in understanding long-run impacts outside of Europe or in places without European or even Indo-European names.

I offer a new context and new data with numeric administrative links to test how string-matching methods compare against the truth sample: data on the entire population of thousands of Indian indentured servants in Fiji from the late nineteenth and early twentieth centuries. I ask, does string matching lead to the same results? If not, can the bias be signed?

I leverage three unique aspects of the data and context. First, individuals received unique numeric identifiers, which lead me to describe the data as "proto-administrative." These identifiers enable me to create a truth sample by linking people over time correctly. Second, I am able to link individuals to two additional datasets created from official records on mortality and return migration. Both datasets were updated continuously by officials and completely covered the population. This means that I neither have to make assumptions about individuals who do not match nor have mechanical gaps in coverage caused by long time periods between data, e.g., in decennial censuses. Third, I examine non-western names with a variety of language origins.

Preliminary results show that string matching fares poorly against two main metrics, mortality and return migration to India. String matching sometimes results in incorrect aggregate statistics on both. More importantly, the composition of deaths and return migrants deviates from the true composition. Finally, the bias cannot be signed in my main binary outcomes, a standard property with non-classical measurement error and varying probabilities of correct matching.