Data simulation scheme

12/15/2017

Goal: Simulate datasets for small area estimation models performed as a post-record linkage step.

Challenge: Need datasets to be realistic and of reasonable size.

1 Proposed method

The idea is to simulate two datasets that are mean to mimic a temporal survey and add duplication and noise to the data.

1. Obtain full empirical conditional histograms from the American Community Survey (ACS).
2. Simulate individuals from the ACS using a Gibbs sampler and the conditionals from 1. Call this datasat D_0.
3. Sample n_1 records from D_0 without replacement.
4. According to some probability of repeat data, p, choose $n_1 * p$ records to occur in the dataset D_1 multiple times.
   - Of these duplicated records, add noise to a prechosen number of fields by simulating those from the empirical distribution (or perhaps according to string distance, depending on the field type).
   - Call this dataset D_1.
5. According to historical immigration patterns, choose a percentage of individuals that would remain in the same area and thus surveyed a second time, q.
6. Include $n_1 * q$ of the original individuals from D_1 (without duplication).
7. Of those records in D_0 not selected for D_1, sample $n_2 - n_1 * q$ records to complete the next dataset.
8. Repeat step 4. to duplicate and distort records and form D_2.

Outcome: Two datasets with distortion and duplication that mimic two years of the ACS survey which can then be used for record linkage and modeling tasks.